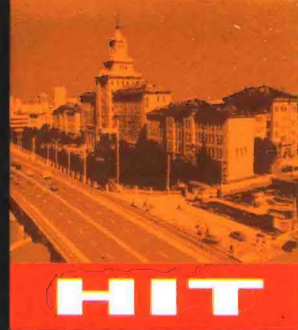




ELSEVIER
爱思唯尔

Encyclopedia of Mathematical Physics (Volume 5)



国外优秀数学著作
原版系列

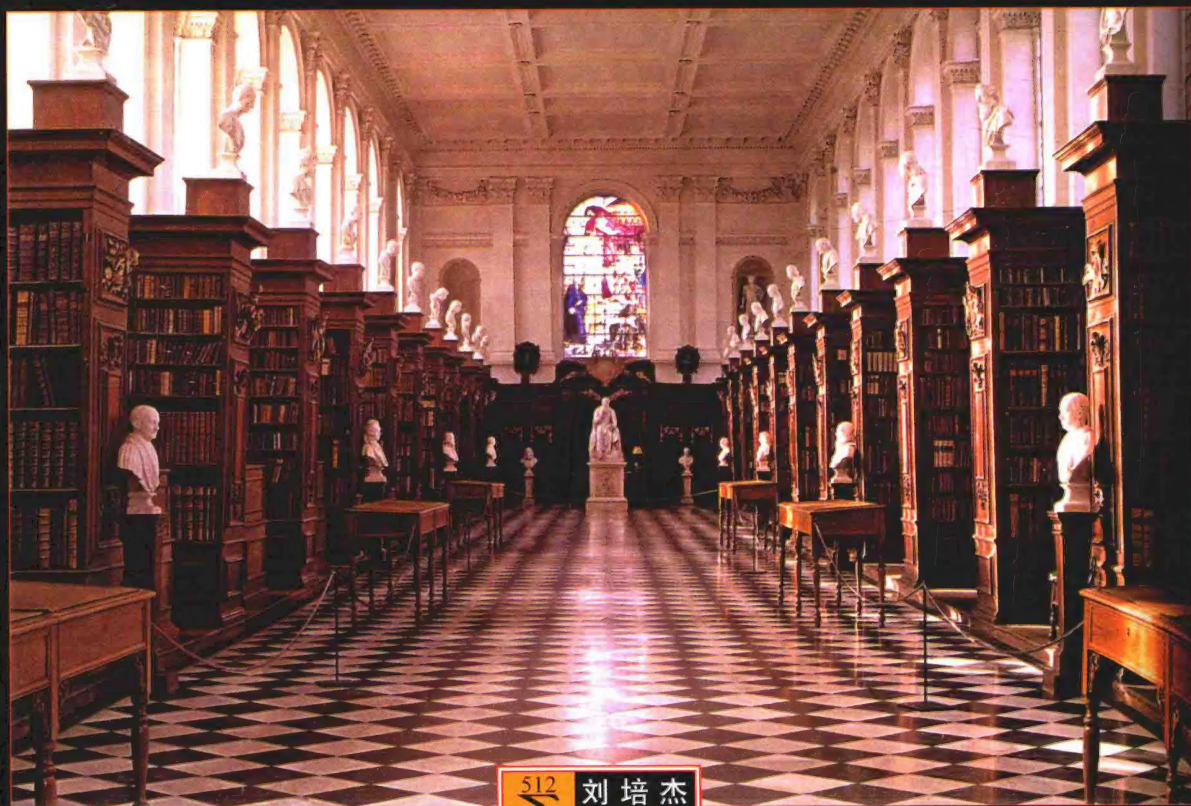
数学物理大百科全书 (第5卷)

Francoise, J. P. (弗朗索瓦斯) Naber, G. L. (纳伯) Tsou Sheung Tsun (孙圣周) 著



哈尔滨工业大学出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

Encyclopedia of Mathematical Physics (Volume 5)



512
 $\sum_{i=0}$ 刘培杰
数学工作室

图片来源:《伟人的建筑师》

数学物理大百科全书 (共5卷)

- 数学物理大百科全书 (第1卷)
- 数学物理大百科全书 (第2卷)
- 数学物理大百科全书 (第3卷)
- 数学物理大百科全书 (第4卷)
- 数学物理大百科全书 (第5卷)

培杰数学国际文化传播中心

www.impj.cn

刘培杰数学工作室网站

<http://lpj.hit.edu.cn>

策划编辑 刘培杰

责任编辑 张永芹 钱辰琛

封面设计 孙茵艾



哈尔滨工业大学出版社 刘培杰数学工作室

联系地址: 哈尔滨市南岗区复华四道街10号

邮编: 150006

联系电话: 0451-86281378 13904613167

E-mail: lpj1378@163.com

微信: impjpp

ISBN 978-7-5603-5525-2



9 787560 355252 >

定价 368.00 元

国外优秀数学著作
原版系列

Encyclopedia of Mathematical Physics (Volume 5)

数学物理大百科全书

(第5卷)

● Francoise, J. P. (弗朗索瓦斯) Naber, G. L. (纳伯) Tsou Sheung Tsun (孙圣周) 著



哈尔滨工业大学出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

黑版贸审字 08-2015-027 号

Encyclopedia of Mathematical Physics

Jean-Pierre Francoise, Gregory L. Naber, and Tsou Sheung Tsun

ISBN:9780125126601

Copyright © 2006 Elsevier Ltd. All rights reserved.

Authorized English language reprint edition published by Elsevier (Singapore) Pte Ltd. and Harbin Institute of Technology Press

Copyright © 2016 by Elsevier (Singapore) Pte Ltd. All rights reserved.

Elsevier (Singapore) Pte Ltd.

3 Killiney Road, #08-01 Winsland House I, Singapore 239519

Tel: (65) 6349-0200

Fax: (65) 6733-1817

First Published 2016

2016 年初版

Printed in China by Harbin Institute of Technology Press under special arrangement with Elsevier (Singapore) Pte Ltd. This edition is authorized for sale in China only, excluding Hong Kong SAR, Macao SAR and Taiwan. Unauthorized export of this edition is a violation of the Copyright Act. Violation of this Law is subject to Civil and Criminal Penalties.

本书英文影印版由 Elsevier (Singapore) Pte Ltd. 授权哈尔滨工业大学出版社在中国大陆境内独家发行。本版仅限在中国境内(不包括香港、澳门以及台湾)出版及标价销售。未经许可之出口,视为违反著作权法,将受民事及刑事法律之制裁。

本书封底贴有 Elsevier 防伪标签,无标签者不得销售。

图书在版编目(CIP)数据

数学物理大百科全书=Encyclopedia of mathematical physics. 第5卷:英文/弗朗索瓦斯(Francoise, J. P.),纳伯(Naber, G. L.),孙圣周著. —哈尔滨:哈尔滨工业大学出版社, 2016.1

ISBN 978-7-5603-5525-2

I. ①数… II. ①弗…②纳…③孙… III. ①数学-百科全书-英文②物理学-百科全书-英文 IV. ①O-61

中国版本图书馆 CIP 数据核字(2015)第 185727 号

策划编辑 刘培杰

责任编辑 张永芹 钱辰琛

封面设计 孙茵艾

出版发行 哈尔滨工业大学出版社

社 址 哈尔滨市南岗区复华四道街 10 号 邮编 150006

传 真 0451-86414749

网 址 <http://hitpress.hit.edu.cn>

印 刷 牡丹江邮电印务有限公司

开 本 880mm×1230mm 1/16 印张 38.75 字数 1220 千字

版 次 2016 年 1 月第 1 版 2016 年 1 月第 1 次印刷

书 号 ISBN 978-7-5603-5525-2

定 价 368.00 元

(如因印装质量问题影响阅读,我社负责调换)

EDITORS AND EDITORIAL ADVISORY BOARD

EDITORS

Jean-Pierre Franoise
Universit  P.-M. Curie, Paris VI
Paris, France

Gregory L. Naber
Drexel University
Philadelphia, PA, USA

Tsou Sheung Tsun
University of Oxford
Oxford, UK

EDITORIAL ADVISORY BOARD

Sergio Albeverio
Rheinische Friedrich-Wilhelms-Universit t Bonn
Bonn, Germany

Huzihiro Araki
Kyoto University
Kyoto, Japan

Abhay Ashtekar
Pennsylvania State University
University Park, PA, USA

Andrea Braides
Universit  di Roma "Tor Vergata"
Roma, Italy

Francesco Calogero
Universit  di Roma "La Sapienza"
Roma, Italy

Cecile DeWitt-Morette
The University of Texas at Austin
Austin, TX, USA

Artur Ekert
University of Cambridge
Cambridge, UK

Giovanni Gallavotti
Universit  di Roma "La Sapienza"
Roma, Italy

Simon Gindikin
Rutgers University
Piscataway, NJ, USA

Gennadi Henkin
Universit  P.-M. Curie, Paris VI
Paris, France

Allen C. Hirshfeld
Universit t Dortmund
Dortmund, Germany

Lisa Jeffrey
University of Toronto
Toronto, Canada

T.W.B. Kibble
Imperial College of Science, Technology and Medicine
London, UK

Antti Kupiainen
University of Helsinki
Helsinki, Finland

Shahn Majid
Queen Mary, University of London
London, UK

Barry M. McCoy
State University of New York Stony Brook
Stony Brook, NY, USA

Hiroshi Ooguri
California Institute of Technology
Pasadena, CA, USA

Roger Penrose
University of Oxford
Oxford, UK

Pierre Ramond
University of Florida
Gainesville, FL, USA

Tudor Ratiu
Ecole Polytechnique Federale de Lausanne
Lausanne, Switzerland

Rudolf Schmid
Emory University
Atlanta, GA, USA

Albert Schwarz
University of California
Davis, CA, USA

Yakov Sinai
Princeton University
Princeton, NJ, USA

Herbert Spohn
Technische Universität München
München, Germany

Stephen J. Summers
University of Florida
Gainesville, FL, USA

Roger Temam
Indiana University
Bloomington, IN, USA

Craig A. Tracy
University of California
Davis, CA, USA

Andrzej Trautman
Warsaw University
Warsaw, Poland

Vladimir Turaev
Institut de Recherche Mathématique Avancée,
Strasbourg, France

Gabriele Veneziano
CERN, Genève, Switzerland

Reinhard F. Werner
Technische Universität Braunschweig
Braunschweig, Germany

C.N. Yang
Tsinghua University
Beijing, China

Eberhard Zeidler
Max-Planck Institut für Mathematik in
den Naturwissenschaften
Leipzig, Germany

Steve Zelditch
Johns Hopkins University
Baltimore, MD, USA

FOREWORD

In bygone centuries, our physical world appeared to be filled to the brim with mysteries. Divine powers could provide for genuine miracles; water and sunlight could turn arid land into fertile pastures, but the same powers could lead to miseries and disasters. The force of life, the *vis vitalis*, was assumed to be the special agent responsible for all living things. The heavens, whatever they were for, contained stars and other heavenly bodies that were the exclusive domain of the Gods.

Mathematics did exist, of course. Indeed, there was one aspect of our physical world that was recognised to be controlled by precise, mathematical logic: the geometric structure of space, elaborated to become a genuine form of art by the ancient Greeks. From my perspective, the Greeks were the first practitioners of ‘mathematical physics’, when they discovered that all geometric features of space could be reduced to a small number of axioms. Today, these would be called ‘fundamental laws of physics’. The fact that the flow of *time* could be addressed with similar exactitude, and that it could be handled geometrically together with space, was only recognised much later. And, yes, there were a few crazy people who were interested in the magic of numbers, but the real world around us seemed to contain so much more that was way beyond our capacities of analysis.

Gradually, all this changed. The Moon and the planets appeared to follow geometrical laws. Galilei and Newton managed to identify their logical rules of motion, and by noting that the concept of mass could be applied to things in the sky just like apples and cannon balls on Earth, they made the sky a little bit more accessible to us. Electricity, magnetism, light and sound were also found to behave in complete accordance with mathematical equations.

Yet all of this was just a beginning. The real changes came with the twentieth century. A completely new way of thinking, by emphasizing mathematical, logical analysis rather than empirical evidence, was pioneered by Albert Einstein. Applying advanced mathematical concepts, only known to a few pure mathematicians, to notions as mundane as space and time, was new to the physicists of his time. Einstein himself had a hard time struggling through the logic of connections and curvatures, notions that were totally new to him, but are only too familiar to students of mathematical physics today. Indeed, there is no better testimony of Einstein’s deep insights at that time, than the fact that we now teach these things regularly in our university classrooms.

Special and general relativity are only small corners of the realm of modern physics that is presently being studied using advanced mathematical methods. We have notoriously complex subjects such as phase transitions in condensed matter physics, superconductivity, Bose–Einstein condensation, the quantum Hall effect, particularly the fractional quantum Hall effect, and numerous topics from elementary particle physics, ranging from fibre bundles and renormalization groups to supergravity, algebraic topology, superstring theory, Calabi–Yau spaces and what not, all of which require the utmost of our mental skills to comprehend them.

The most bewildering observation that we make today is that it seems that our *entire* physical world appears to be controlled by mathematical equations, and these are not just sloppy and debatable models, but precisely documented properties of materials, of systems, and of phenomena in all echelons of our universe.

Does this really apply to our entire world, or only to parts of it? Do features, notions, entities exist that are emphatically *not* mathematical? What about intuition, or dreams, and what about consciousness? What about religion? Here, most of us would say, one should not even try to apply mathematical analysis, although even here, some brave social scientists are making attempts at coordinating rational approaches.

No, there are clear and important differences between the physical world and the mathematical world. Where the physical world stands out is the fact that it refers to ‘reality’, whatever ‘reality’ is. Mathematics is the world of pure logic and pure reasoning. In physics, it is the experimental evidence that ultimately decides whether a theory is acceptable or not. Also, the methodology in physics is different.

A beautiful example is the serendipitous discovery of superconductivity. In 1911, the Dutch physicist Heike Kamerlingh Onnes was the first to achieve the liquefaction of helium, for which a temperature below 4.25 K had to be realized. Heike decided to measure the specific conductivity of mercury, a metal that is frozen solid at such low temperatures. But something appeared to go wrong during the measurements, since the volt meter did not show any voltage at all. All experienced physicists in the team assumed that they were dealing with a malfunction. It would not have been the first time for a short circuit to occur in the electrical equipment, but, this time, in spite of several efforts, they failed to locate it. One of the assistants was responsible for keeping the temperature of the sample well within that of liquid helium, a dull job, requiring nothing else than continuously watching some dials. During one of the many tests, however, he dozed off. The temperature rose, and suddenly the measurements showed the normal values again. It then occurred to the investigators that the effect and its temperature dependence were completely reproducible. Below 4.19 degrees Kelvin the conductivity of mercury appeared to be strictly infinite. Above that temperature, it is finite, and the transition is a very sudden one. Superconductivity was discovered (D. van Delft, “Heike Kamerling Onnes”, Uitgeverij Bert Bakker, Amsterdam, 2005 (in Dutch)).

This is not the way mathematical discoveries are made. Theorems are not produced by assistants falling asleep, even if examples do exist of incidents involving some miraculous fortune.

The hybrid science of mathematical physics is a very curious one. Some of the topics in this Encyclopedia are undoubtedly physical. High T_c superconductivity, breaking water waves, and magneto-hydrodynamics, are definitely topics of physics where experimental data are considered more decisive than any high-brow theory. Cohomology theory, Donaldson–Witten theory, and AdS/CFT correspondence, however, are examples of purely mathematical exercises, even if these subjects, like all of the others in this compilation, are strongly inspired by, and related to, questions posed in physics.

It is inevitable, in a compilation of a large number of short articles with many different authors, to see quite a bit of variation in style and level. In this Encyclopedia, theoretical physicists as well as mathematicians together made a huge effort to present in a concise and understandable manner their vision on numerous important issues in advanced mathematical physics. All include references for further reading. We hope and expect that these efforts will serve a good purpose.

Gerard 't Hooft,
Spinoza Institute,
Utrecht University,
The Netherlands.

PREFACE

Mathematical Physics as a distinct discipline is relatively new. The International Association of Mathematical Physics was founded only in 1976. The interaction between physics and mathematics has, of course, existed since ancient times, but the recent decades, perhaps partly because we are living through them, appear to have witnessed tremendous progress, yielding new results and insights at a dizzying pace, so much so that an encyclopedia seems now needed to collate the gathered knowledge.

Mathematical Physics brings together the two great disciplines of Mathematics and Physics to the benefit of both, the relationship between them being symbiotic. On the one hand, it uses mathematics as a tool to organize physical ideas of increasing precision and complexity, and on the other it draws on the questions that physicists pose as a source of inspiration to mathematicians. A classical example of this relationship exists in Einstein's theory of relativity, where differential geometry played an essential role in the formulation of the physical theory while the problems raised by the ensuing physics have in turn boosted the development of differential geometry. It is indeed a happy coincidence that we are writing now a preface to an encyclopedia of mathematical physics in the centenary of Einstein's *annus mirabilis*.

The project of putting together an encyclopedia of mathematical physics looked, and still looks, to us a formidable enterprise. We would never have had the courage to undertake such a task if we did not believe, first, that it is worthwhile and of benefit to the community, and second, that we would get the much-needed support from our colleagues. And this support we did get, in the form of advice, encouragement, and practical help too, from members of our Editorial Advisory Board, from our authors, and from others as well, who have given unstintingly so much of their time to help us shape this Encyclopedia.

Mathematical Physics being a relatively new subject, it is not yet clearly delineated and could mean different things to different people. In our choice of topics, we were guided in part by the programs of recent International Congresses on Mathematical Physics, but mainly by the advice from our Editorial Advisory Board and from our authors. The limitations of space and time, as well as our own limitations, necessitated the omission of certain topics, but we have tried to include all that we believe to be core subjects and to cover as much as possible the most active areas.

Our subject being interdisciplinary, we think it appropriate that the Encyclopedia should have certain special features. Applications of the same mathematical theory, for instance, to different problems in physics will have different emphasis and treatment. By the same token, the same problem in physics can draw upon resources from different mathematical fields. This is why we divide the Encyclopedia into two broad sections: physics subjects and related mathematical subjects. Articles in either section are deliberately allowed a fair amount of overlap with one another and many articles will appear under more than one heading, but all are linked together by elaborate cross referencing. We think this gives a better picture of the subject as a whole and will serve better a community of researchers from widely scattered yet related fields.

The Encyclopedia is intended primarily for experienced researchers but should be of use also to beginning graduate students. For the latter category of readers, we have included eight elementary introductory articles for easy reference, with those on mathematics aimed at physics graduates and those on physics aimed at mathematics graduates, so that these articles can serve as their first port of call to enable them to embark on any of the main articles without the need to consult other material beforehand. In fact, we think these articles may even form the

foundation of advanced undergraduate courses, as we know that some authors have already made such use of them.

In addition to the printed version, an on-line version of the Encyclopedia is planned, which will allow both the contents and the articles themselves to be updated if and when the occasion arises. This is probably a necessary provision in such a rapidly advancing field.

This project was some four years in the making. Our foremost thanks at its completion go to the members of our Editorial Advisory Board, who have advised, helped and encouraged us all along, and to all our authors who have so generously devoted so much of their time to writing these articles and given us much useful advice as well. We ourselves have learnt a lot from these colleagues, and made some wonderful contacts with some among them. Special thanks are due also to Arthur Greenspoon whose technical expertise was indispensable.

The project was started with Academic Press, which was later taken over by Elsevier. We thank warmly members of their staff who have made this transition admirably seamless and gone on to assist us greatly in our task: both Carey Chapman and Anne Guillaume, who were in charge of the whole project and have been with us since the beginning, and Edward Taylor responsible for the copy-editing. And Martin Ruck, who manages to keep an overwhelming amount of details constantly at his fingertips, and who is never known to have lost a single email, deserves a very special mention.

As a postscript, we would like to express our gratitude to the very large number of authors who generously agreed to donate their honorariums to support the Committee for Developing Countries of the European Mathematical Society in their work to help our less fortunate colleagues in the developing world.

Jean-Pierre Franoise
Gregory L. Naber
Tsou Sheung Tsun

GUIDE TO USE OF THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in this Encyclopedia is organised into two sections. At the start of Volume 1 are eight **Introductory Articles**. The introductory articles on mathematics are aimed at physics graduates; those on physics are aimed at mathematics graduates. It is intended that these articles should serve as the first port of call for graduate students, to enable them to embark on any of the main entries without the need to consult other material beforehand.

Following the Introductory Articles, the main body of the Encyclopedia is arranged as a series of entries in alphabetical order. These entries fill the remainder of Volume 1 and all of the subsequent volumes (2–5).

To help you realize the full potential of the material in the Encyclopedia we have provided four features to help you find the topic of your choice: a contents list by subject, an alphabetical contents list, cross-references, and a full subject index.

1. Contents List by Subject

Your first point of reference will probably be the contents list by subject. This list appears at the front of each volume, and groups the entries under subject headings describing the broad themes of mathematical physics. This will enable the reader to make quick connections between entries and to locate the entry of interest. The contents list by subject is divided into two main sections: *Physics Subjects* and *Related Mathematics Subjects*. Under each main section heading, you will find several subject areas (such as GENERAL RELATIVITY in Physics Subjects or NONCOMMUTATIVE GEOMETRY in Related Mathematics Subjects). Under each subject area is a list of those entries that cover aspects of that subject, together with the volume and page numbers on which these entries may be found.

Because mathematical physics is so highly interconnected, individual entries may appear under more than one subject area. For example, the entry GAUGE THEORY: MATHEMATICAL APPLICATIONS is listed under the Physics Subject GAUGE THEORY as well as in a broad range of Related Mathematics Subjects.

2. Alphabetical Contents List

The alphabetical contents list, which also appears at the front of each volume, lists the entries in the order in which they appear in the Encyclopedia. This list provides both the volume number and the page number of the entry.

You will find “dummy entries” where obvious synonyms exist for entries or where we have grouped together related topics. Dummy entries appear in both the contents list and the body of the text.

Example

If you were attempting to locate material on path integral methods via the alphabetical contents list:

PATH INTEGRAL METHODS *see* Functional Integration in Quantum Physics; Feynman Path Integrals

The dummy entry directs you to two other entries in which path integral methods are covered. At the appropriate locations in the contents list, the volume and page numbers for these entries are given.

If you were trying to locate the material by browsing through the text and you had looked up Path Integral Methods, then the following information would be provided in the dummy entry:

Path Integral Methods *see* Functional Integration in Quantum Physics; Feynman Path Integrals

3. Cross-References

All of the articles in the Encyclopedia have been extensively cross-referenced. The cross-references, which appear at the end of an entry, serve three different functions:

- i. To indicate if a topic is discussed in greater detail elsewhere.
- ii. To draw the reader's attention to parallel discussions in other entries.
- iii. To indicate material that broadens the discussion.

Example

The following list of cross-references appears at the end of the entry STOCHASTIC HYDRODYNAMICS

See also: Cauchy Problem for Burgers-Type Equations; Hamiltonian Fluid Dynamics; Incompressible Euler Equations: Mathematical Theory; Malliavin Calculus; Non-Newtonian Fluids; Partial Differential Equations: Some Examples; Stochastic Differential Equations; Turbulence Theories; Viscous Incompressible Fluids: Mathematical Theory; Vortex Dynamics

Here you will find examples of all three functions of the cross-reference list: a topic discussed in greater detail elsewhere (e.g. Incompressible Euler Equations: Mathematical Theory), parallel discussion in other entries (e.g. Stochastic Differential Equations) and reference to entries that broaden the discussion (e.g. Turbulence Theories).

The eight Introductory Articles are not cross-referenced from any of the main entries, as it is expected that introductory articles will be of general interest. As mentioned above, the Introductory Articles may be found at the start of Volume 1.

4. Index

The index will provide you with the volume and page number where the material is located. The index entries differentiate between material that is a whole entry, is part of an entry, or is data presented in a figure or table. Detailed notes are provided on the opening page of the index.

5. Contributors

A full list of contributors appears at the beginning of each volume.

CONTRIBUTORS

A Abbondandolo

Università di Pisa
Pisa, Italy

M J Ablowitz

University of Colorado
Boulder, CO, USA

S L Adler

Institute for Advanced Study
Princeton, NJ, USA

H Airault

Université de Picardie
Amiens, France

G Alberti

Università di Pisa
Pisa, Italy

S Albeverio

Rheinische Friedrich–Wilhelms-Universität Bonn
Bonn, Germany

S T Ali

Concordia University
Montreal, QC, Canada

R Alicki

University of Gdańsk
Gdańsk, Poland

G Altarelli

CERN
Geneva, Switzerland

C Amrouche

Université de Pau et des Pays de l'Adour
Pau, France

M Anderson

State University of New York at Stony Brook
Stony Brook, NY, USA

L Andersson

University of Miami
Coral Gables, FL, USA and Albert Einstein Institute
Potsdam, Germany

B Andreas

Humboldt-Universität zu Berlin
Berlin, Germany

V Araújo

Universidade do Porto
Porto, Portugal

A Ashtekar

Pennsylvania State University
University Park, PA, USA

W Van Assche

Katholieke Universiteit Leuven
Leuven, Belgium

G Aubert

Université de Nice Sophia Antipolis
Nice, France

H Au-Yang

Oklahoma State University
Stillwater, OK, USA

M A Aziz-Alaoui

Université du Havre
Le Havre, France

V Bach

Johannes Gutenberg-Universität
Mainz, Germany

C Bachas

Ecole Normale Supérieure
Paris, France

V Baladi

Institut Mathématique de Jussieu
Paris, France

D Bambusi

Università di Milano
Milan, Italy

C Bardos

Université de Paris 7
Paris, France

D Bar-Natan

University of Toronto
Toronto, ON, Canada

E L Basor

California Polytechnic State University
San Luis Obispo, CA, USA

M T Batchelor

Australian National University
Canberra, ACT, Australia

S Bauer

Universität Bielefeld
Bielefeld, Germany

V Beffara

Ecole Normale Supérieure de Lyon
Lyon, France

R Beig

Universität Wien
Vienna, Austria

M I Belishev

Petersburg Department of Steklov Institute
of Mathematics
St. Petersburg, Russia

P Bernard

Université de Paris Dauphine
Paris, France

D Birmingham

University of the Pacific
Stockton, CA, USA

Jiří Bičák

Charles University, Prague, Czech Republic
and Albert Einstein Institute
Potsdam, Germany

C Blanchet

Université de Bretagne-Sud
Vannes, France

M Blasone

Università degli Studi di Salerno
Baronissi (SA), Italy

M Blau

Université de Neuchâtel
Neuchâtel, Switzerland

S Boatto

IMPA
Rio de Janeiro, Brazil

L V Bogachev

University of Leeds
Leeds, UK

L Boi

EHESS and LUTH
Paris, France

M Bojowald

The Pennsylvania State University
University Park, PA, USA

C Bonatti

Université de Bourgogne
Dijon, France

P Bonckaert

Universiteit Hasselt
Diepenbeek, Belgium

F Bonetto

Georgia Institute of Technology
Atlanta, GA, USA

G Bouchitté

Université de Toulon et du Var
La Garde, France

A Bovier

Weierstrass Institute for Applied Analysis and Stochastics
Berlin, Germany

H W Braden

University of Edinburgh
Edinburgh, UK

H Bray

Duke University
Durham, NC, USA

Y Brenier

Université de Nice Sophia Antipolis
Nice, France

J Bros
CEA/DSM/SPhT, CEA/Saclay
Gif-sur-Yvette, France

R Brunetti
Universität Hamburg
Hamburg, Germany

M Bruschi
Università di Roma “La Sapienza”
Rome, Italy

T Brzeziński
University of Wales Swansea
Swansea, UK

D Buchholz
Universität Göttingen
Göttingen, Germany

N Burq
Université Paris-Sud
Orsay, France

F H Busse
Universität Bayreuth
Bayreuth, Germany

G Buttazzo
Università di Pisa
Pisa, Italy

P Buttà
Università di Roma “La Sapienza”
Rome, Italy

S L Cacciatori
Università di Milano
Milan, Italy

P T Callaghan
Victoria University of Wellington
Wellington, New Zealand

Francesco Calogero
University of Rome, Rome, Italy and Institute
Nazionale di Fisica Nucleare
Rome, Italy

A Carati
Università di Milano
Milan, Italy

J Cardy
Rudolf Peierls Centre for Theoretical Physics
Oxford, UK

R Caseiro
Universidade de Coimbra
Coimbra, Portugal

A S Cattaneo
Universität Zürich
Zürich, Switzerland

A Celletti
Università di Roma “Tor Vergata”
Rome, Italy

D Chae
Sungkyunkwan University
Suwon, South Korea

G-Q Chen
Northwestern University
Evanston, IL, USA

L Chierchia
Università degli Studi “Roma Tre”
Rome, Italy

S Chmutov
Petersburg Department of Steklov
Institute of Mathematics
St. Petersburg, Russia

M W Choptuik
University of British Columbia
Vancouver, Canada

Y Choquet-Bruhat
Université P.-M. Curie, Paris VI
Paris, France

P T Chruściel
Université de Tours
Tours, France

Chong-Sun Chu
University of Durham
Durham, UK

F Cipriani
Politecnico di Milano
Milan, Italy

R L Cohen

Stanford University
Stanford, CA, USA

T H Colding

University of New York
New York, NY, USA

J C Collins

Penn State University
University Park, PA, USA

G Comte

Université de Nice Sophia Antipolis
Nice, France

A Constantin

Trinity College
Dublin, Republic of Ireland

D Crowdy

Imperial College
London, UK

A B Cruzeiro

University of Lisbon
Lisbon, Portugal

G Dal Maso

SISSA
Trieste, Italy

F Dalfovo

Università di Trento
Povo, Italy

A S Dancer

University of Oxford
Oxford, UK

P D’Ancona

Università di Roma “La Sapienza”
Rome, Italy

S R Das

University of Kentucky
Lexington, KY, USA

E Date

Osaka University
Osaka, Japan

N Datta

University of Cambridge
Cambridge, UK

G W Delius

University of York
York, UK

G F dell’Antonio

Università di Roma “La Sapienza”
Rome, Italy

C DeWitt-Morette

The University of Texas at Austin
Austin, TX, USA

L Diósi

Research Institute for Particle and Nuclear Physics
Budapest, Hungary

A Doliwa

University of Warmia and Mazury in Olsztyn
Olsztyn, Poland

G Dolzmann

University of Maryland
College Park, MD, USA

S K Donaldson

Imperial College
London, UK

T C Dorlas

Dublin Institute for Advanced Studies
Dublin, Republic of Ireland

M R Douglas

Rutgers, The State University of New Jersey
Piscataway, NJ, USA

M Dütsch

Universität Zürich
Zürich, Switzerland

B Dubrovin

SISSA-ISAS
Trieste, Italy

J J Duistermaat

Universiteit Utrecht
Utrecht, The Netherlands

S Duzhin

Petersburg Department of Steklov Institute of
Mathematics
St. Petersburg, Russia

G Ecker
Universität Wien
Vienna, Austria

M Efendiev
Universität Stuttgart
Stuttgart, Germany

T Eguchi
University of Tokyo
Tokyo, Japan

J Ehlers
Max Planck Institut für Gravitationsphysik
(Albert-Einstein Institut)
Golm, Germany

P E Ehrlich
University of Florida
Gainesville, FL, USA

D Einzel
Bayerische Akademie der Wissenschaften
Garching, Germany

G A Elliott
University of Toronto
Toronto, Canada

G F R Ellis
University of Cape Town
Cape Town, South Africa

C L Epstein
University of Pennsylvania
Philadelphia, PA, USA

J Escher
Universität Hannover
Hannover, Germany

J B Etnyre
University of Pennsylvania
Philadelphia, PA, USA

G Falkovich
Weizmann Institute of Science
Rehovot, Israel

M Farge
Ecole Normale Supérieure
Paris, France

B Ferrario
Università di Pavia
Pavia, Italy

R Finn
Stanford University
Stanford, CA, USA

D Fiorenza
Università di Roma “La Sapienza”
Rome, Italy

A E Fischer
University of California
Santa Cruz, CA, USA

A S Fokas
University of Cambridge
Cambridge, UK

J-P Françoise
Université P.-M. Curie, Paris VI
Paris, France

S Franz
The Abdus Salam ICTP
Trieste, Italy

L Frappat
Université de Savoie
Chambery-Annecy, France

J Frauendiener
Universität Tübingen
Tübingen, Germany

K Fredenhagen
Universität Hamburg
Hamburg, Germany

S Friedlander
University of Illinois-Chicago
Chicago, IL, USA

M R Gaberdiel
ETH Zürich
Zürich, Switzerland

G Gaeta
Università di Milano
Milan, Italy

L Galgani Università di Milano Milan, Italy	H Gottschalk Rheinische Friedrich-Wilhelms-Universität Bonn Bonn, Germany
G Gallavotti Università di Roma “La Sapienza” Rome, Italy	O Goubet Université de Picardie Jules Verne Amiens, France
R Gambini Universidad de la República Montevideo, Uruguay	T R Govindarajan The Institute of Mathematical Sciences Chennai, India
G Gentile Università degli Studi “Roma Tre” Rome, Italy	A Grassi University of Pennsylvania Philadelphia, PA, USA
A Di Giacomo Università di Pisa Pisa, Italy	P G Grinevich L D Landau Institute for Theoretical Physics Moscow, Russia
P B Gilkey University of Oregon Eugene, OR, USA	Ch Gruber Ecole Polytechnique Fédérale de Lausanne Lausanne, Switzerland
R Gilmore Drexel University Philadelphia, PA, USA	J-L Guermond Université de Paris Sud Orsay, France
S Gindikin Rutgers University Piscataway, NJ, USA	F Guerra Università di Roma “La Sapienza” Rome, Italy
A Giorgilli Università di Milano Milan, Italy	T Guhr Lunds Universitet Lund, Sweden
G A Goldin Rutgers University Piscataway, NJ, USA	C Guillopé Université Paris XII – Val de Marne Créteil, France
G González Louisiana State University Baton Rouge, LA, USA	C Gundlach University of Southampton Southampton, UK
R Gopakumar Harish-Chandra Research Institute Allahabad, India	S Gutt Université Libre de Bruxelles Brussels, Belgium
D Gottesman Perimeter Institute Waterloo, ON, Canada	K Hannabuss University of Oxford Oxford, UK

M Haragus

Université de Franche-Comté
Besançon, France

S G Harris

St. Louis University
St. Louis, MO, USA

B Hasselblatt

Tufts University
Medford, MA, USA

P Hayden

McGill University
Montreal, QC, Canada

D C Heggie

The University of Edinburgh
Edinburgh, UK

B Helffer

Université Paris-Sud
Orsay, France

G M Henkin

Université P.-M. Curie, Paris VI
Paris, France

M Henneaux

Université Libre de Bruxelles
Bruxelles, Belgium

S Herrmann

Université Henri Poincaré, Nancy 1
Vandoeuvre-lès-Nancy, France

C P Herzog

University of California at Santa Barbara
Santa Barbara, CA, USA

J G Heywood

University of British Columbia
Vancouver, BC, Canada

A C Hirshfeld

Universität Dortmund
Dortmund, Germany

A S Holevo

Steklov Mathematical Institute
Moscow, Russia

T J Hollowood

University of Wales Swansea
Swansea, UK

D D Holm

Imperial College
London, UK

J-W van Holten

NIKHEF
Amsterdam, The Netherlands

A Huckleberry

Ruhr-Universität Bochum
Bochum, Germany

K Hulek

Universität Hannover
Hannover, Germany

D Iagolnitzer

CEA/DSM/SPHT, CEA/Saclay
Gif-sur-Yvette, France

R Illge

Friedrich-Schiller-Universität Jena
Jena, Germany

P Imkeller

Humboldt Universität zu Berlin
Berlin, Germany

G Iooss

Institut Non Linéaire de Nice
Valbonne, France

M Irigoyen

Université P.-M. Curie, Paris VI
Paris, France

J Isenberg

University of Oregon
Eugene, OR, USA

R Ivanova

University of Hawaii Hilo
Hilo, HI, USA

E M Izhikevich

The Neurosciences Institute
San Diego, CA, USA

R W Jackiw

Massachusetts Institute of Technology
Cambridge, MA, USA

J K Jain

The Pennsylvania State University
University Park, PA, USA

M Jardim

IMECC–UNICAMP
Campinas, Brazil

L C Jeffrey

University of Toronto
Toronto, ON, Canada

J Jiménez

Universidad Politecnica de Madrid
Madrid, Spain

S Jitomirskaya

University of California at Irvine
Irvine, CA, USA

P Jizba

Czech Technical University
Prague, Czech Republic

A Joets

Université Paris-Sud
Orsay, France

K Johansson

Kungl Tekniska Högskolan
Stockholm, Sweden

G Jona-Lasinio

Università di Roma “La Sapienza”
Rome, Italy

V F R Jones

University of California at Berkeley
Berkeley, CA, USA

N Joshi

University of Sydney
Sydney, NSW, Australia

D D Joyce

University of Oxford
Oxford, UK

C D Jäkel

Ludwig-Maximilians-Universität München
München, Germany

G Kasperski

Université Paris-Sud XI
Orsay, France

L H Kauffman

University of Illinois at Chicago
Chicago, IL, USA

R K Kaul

The Institute of Mathematical Sciences
Chennai, India

Y Kawahigashi

University of Tokyo
Tokyo, Japan

B S Kay

University of York
York, UK

R Kenyon

University of British Columbia
Vancouver, BC, Canada

M Keyl

Università di Pavia
Pavia, Italy

T W B Kibble

Imperial College
London, UK

S Kichenassamy

Université de Reims Champagne-Ardenne
Reims, France

J Kim

University of California at Irvine
Irvine, USA

S B Kim

Chonnam National University
Gwangju, South Korea

A Kirillov

University of Pennsylvania
Philadelphia, PA, USA

A Kirillov, Jr.

Stony Brook University
Stony Brook, NY, USA

K Kirsten

Baylor University
Waco, TX, USA

F Kirwan University of Oxford Oxford, UK	M Krbec Academy of Sciences Prague, Czech Republic
S Klainerman Princeton University Princeton, NJ, USA	D Kreimer IHES Bures-sur-Yvette, France
I R Klebanov Princeton University Princeton, NJ, USA	A Kresch University of Warwick Coventry, UK
Y Kondratiev Universität Bielefeld Bielefeld, Germany	D Kretschmann Technische Universität Braunschweig Braunschweig, Germany
A Konechny Rutgers, The State University of New Jersey Piscataway, NJ, USA	P B Kronheimer Harvard University Cambridge, MA, USA
K Konishi Università di Pisa Pisa, Italy	B Kuckert Universität Hamburg Hamburg, Germany
T H Koornwinder University of Amsterdam Amsterdam, The Netherlands	Y Kuramoto Hokkaido University Sapporo, Japan
P Kornprobst INRIA Sophia Antipolis, France	J M F Labastida CSIC Madrid, Spain
V P Kostov Université de Nice Sophia Antipolis Nice, France	G Labrosse Université Paris-Sud XI Orsay, France
R Kotecký Charles University Prague, Czech Republic and the University of Warwick, UK	C Landim IMPA, Rio de Janeiro, Brazil and UMR 6085 and Université de Rouen France
Y Kozitsky Uniwersytet Marii Curie-Sklodowskiej Lublin, Poland	E Langmann KTH Physics Stockholm, Sweden
P Kramer Universität Tübingen Tübingen, Germany	S Laporta Università di Parma Parma, Italy
C Krattenthaler Universität Wien Vienna, Austria	O D Lavrentovich Kent State University Kent, OH, USA

G F Lawler

Cornell University
Ithaca, NY, USA

C Le Bris

CERMICS – ENPC
Champs Sur Marne, France

A Lesne

Université P.-M. Curie, Paris VI
Paris, France

D Levi

Università “Roma Tre”
Rome, Italy

J Lewandowski

Uniwersyte Warszawski
Warsaw, Poland

R G Littlejohn

University of California at Berkeley
Berkeley, CA, USA

R Livi

Università di Firenze
Sesto Fiorentino, Italy

R Longoni

Università di Roma “La Sapienza”
Rome, Italy

J Lowengrub

University of California at Irvine
Irvine, USA

C Lozano

INTA
Torrejón de Ardoz, Spain

T T Q Lê

Georgia Institute of Technology
Atlanta, GA, USA

B Lucquin-Desreux

Université P.-M. Curie, Paris VI
Paris, France

V Lyubashenko

Institute of Mathematics
Kyiv, Ukraine

M Lyubich

University of Toronto
Toronto, ON, Canada and Stony Brook University
NY, USA

R Léandre

Université de Bourgogne
Dijon, France

P Lévy

Budapest University of Technology and Economics
Budapest, Hungary

R Maartens

Portsmouth University
Portsmouth, UK

N MacKay

University of York
York, UK

J Magnen

Ecole Polytechnique
France

F Magri

Università di Milano Bicocca
Milan, Italy

J Maharana

Institute of Physics
Bhubaneswar, India

S Majid

Queen Mary, University of London
London, UK

C Marchioro

Università di Roma “La Sapienza”
Rome, Italy

K Marciniak

Linköping University
Norrköping, Sweden

M Marcolli

Max-Planck-Institut für Mathematik
Bonn, Germany

M Mariño

CERN
Geneva, Switzerland

J Marklof

University of Bristol
Bristol, UK

C-M Marle

Université P.-M. Curie, Paris VI
Paris, France

L Mason

University of Oxford
Oxford, UK

V Mastropietro

Università di Roma “Tor Vergata”
Rome, Italy

V Mathai

University of Adelaide
Adelaide, SA, Australia

J Mawhin

Université Catholique de Louvain
Louvain-la-Neuve, Belgium

S Mazzucchi

Università di Trento
Povo, Italy

B M McCoy

State University of New York at Stony Brook
Stony Brook, NY, USA

E Meinrenken

University of Toronto
Toronto, ON, Canada

I Melbourne

University of Surrey
Guildford, UK

J Mickelsson

KTH Physics
Stockholm, Sweden

W P Minicozzi II

University of New York
New York, NY, USA

S Miracle-Solé

Centre de Physique Théorique, CNRS
Marseille, France

A Miranville

Université de Poitiers
Chasseneuil, France

P K Mitter

Université de Montpellier 2
Montpellier, France

V Moncrief

Yale University
New Haven, CT, USA

Y Morita

Ryukoku University
Otsu, Japan

P J Morrison

University of Texas at Austin
Austin, TX, USA

J Mund

Universidade de São Paulo
São Paulo, Brazil

F Musso

Università “Roma Tre”
Rome, Italy

G L Naber

Drexel University
Philadelphia, PA, USA

B Nachtergaele

University of California at Davis
Davis, CA, USA

C Nash

National University of Ireland
Maynooth, Ireland

Š Nečasová

Academy of Sciences
Prague, Czech Republic

A I Neishtadt

Russian Academy of Sciences
Moscow, Russia

N Neumaier

Albert-Ludwigs-University in Freiburg
Freiburg, Germany

S E Newhouse

Michigan State University
E. Lansing, MI, USA

C M Newman

New York University
New York, NY, USA

S Nikčević

SANU
Belgrade, Serbia and Montenegro

M Nitsche

University of New Mexico
Albuquerque, NM, USA

R G Novikov

Université de Nantes
Nantes, France

J M Nunes da Costa

Universidade de Coimbra
Coimbra, Portugal

S O'Brien

Tyndall National Institute
Cork, Republic of Ireland

A Okounkov

Princeton University
Princeton, NJ, USA

A Onuki

Kyoto University
Kyoto, Japan

J-P Ortega

Université de Franche-Comté
Besançon, France

H Osborn

University of Cambridge
Cambridge, UK

Maciej P Wojtkowski

University of Arizona
Tucson, AZ, USA and Institute of Mathematics PAN
Warsaw, Poland

J Palmer

University of Arizona
Tucson, AZ, USA

J H Park

Sungkyunkwan University
Suwon, South Korea

P E Parker

Wichita State University
Wichita KS, USA

S Paycha

Université Blaise Pascal
Aubière, France

P A Pearce

University of Melbourne
Parkville VIC, Australia

P Pearle

Hamilton College
Clinton, NY, USA

M Pedroni

Università di Bergamo
Dalmine (BG), Italy

B Pelloni

University of Reading
UK

R Penrose

University of Oxford
Oxford, UK

A Perez

Penn State University,
University Park, PA, USA

J H H Perk

Oklahoma State University
Stillwater, OK, USA

T Peternell

Universität Bayreuth
Bayreuth, Germany

D Petz

Budapest University of Technology and Economics
Budapest, Hungary

M J Pflaum

Johann Wolfgang Goethe-Universität
Frankfurt, Germany

B Piccoli

Istituto per le Applicazioni del Calcolo
Rome, Italy

C Piquet

Université P.-M. Curie, Paris VI
Paris, France

L P Pitaevskii
Università di Trento
Povo, Italy

S Pokorski
Warsaw University
Warsaw, Poland

E Presutti
Università di Roma “Tor Vergata”
Rome, Italy

E Previato
Boston University
Boston, MA, USA

B Prinari
Università degli Studi di Lecce
Lecce, Italy

J Pullin
Louisiana State University
Baton Rouge, LA, USA

M Pulvirenti
Università di Roma “La Sapienza”
Rome, Italy

O Ragnisco
Università “Roma Tre”
Rome, Italy

P Ramadevi
Indian Institute of Technology Bombay
Mumbai, India

S A Ramakrishna
Indian Institute of Technology
Kanpur, India

J Rasmussen
Princeton University
Princeton, NJ, USA

L Rastelli
Princeton University
Princeton, NJ, USA

T S Ratiu
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

S Rauch-Wojciechowski
Linköping University
Linköping, Sweden

K-H Rehren
Universität Göttingen
Göttingen, Germany

E Remiddi
Università di Bologna
Bologna, Italy

J E Roberts
Università di Roma “Tor Vergata”
Rome, Italy

L Rey-Bellet
University of Massachusetts
Amherst, MA, USA

R Robert
Université Joseph Fourier
Saint Martin D'Hères, France

F A Rogers
King's College London
London, UK

R M S Rosa
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil

C Rovelli
Université de la Méditerranée et Centre
de Physique Théorique
Marseilles, France

S N M Ruijsenaars
Centre for Mathematics and Computer Science
Amsterdam, The Netherlands

F Russo
Université Paris 13
Villetaneuse, France

L H Ryder
University of Kent
Canterbury, UK

S Sachdev
Yale University
New Haven, CT, USA

H Sahlmann
Universiteit Utrecht
Utrecht, The Netherlands

- M Salmhofer**
Universität Leipzig
Leipzig, Germany
- P M Santini**
Università di Roma “La Sapienza”
Rome, Italy
- A Sarmiento**
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
- R Sasaki**
Kyoto University
Kyoto, Japan
- A Savage**
University of Toronto
Toronto, ON, Canada
- M Schechter**
University of California at Irvine
Irvine, CA, USA
- D-M Schlingemann**
Technical University of Braunschweig
Braunschweig, Germany
- R Schmid**
Emory University
Atlanta, GA, USA
- G Schneider**
Universität Karlsruhe
Karlsruhe, Germany
- K Schneider**
Université de Provence
Marseille, France
- B Schroer**
Freie Universität Berlin
Berlin, Germany
- T Schücker**
Université de Marseille
Marseille, France
- S Scott**
King’s College London
London, UK
- P Selick**
University of Toronto
Toronto, ON, Canada
- M A Semenov-Tian-Shansky**
Steklov Institute of Mathematics
St. Petersburg, Russia and and Université de Bourgogne
Dijon, France
- A N Sengupta**
Louisiana State University
Baton Rouge LA, USA
- S Serfaty**
New York University
New York, NY, USA
- E R Sharpe**
University of Utah
Salt Lake City, UT, USA
- D Shepelsky**
Institute for Low Temperature Physics and Engineering
Kharkov, Ukraine
- S Shlosman**
Université de Marseille
Marseille, France
- A Siconolfi**
Università di Roma “La Sapienza”
Rome, Italy
- V Sidoravicius**
IMPA
Rio de Janeiro, Brazil
- J A Smoller**
University of Michigan
Ann Arbor MI, USA
- M Socolovsky**
Universidad Nacional Autónoma de México
México DF, México
- J P Solovej**
University of Copenhagen
Copenhagen, Denmark
- A Soshnikov**
University of California at Davis
Davis, CA, USA
- J M Speight**
University of Leeds
Leeds, UK

H Spohn
Technische Universität München
Garching, Germany

J Stasheff
Lansdale, PA, USA

D L Stein
University of Arizona
Tucson, AZ, USA

K S Stelle
Imperial College
London, UK

G Sterman
Stony Brook University
Stony Brook, NY, USA

S Stringari
Università di Trento
Povo, Italy

S J Summers
University of Florida
Gainesville, FL, USA

V S Sunder
The Institute of Mathematical Sciences
Chennai, India

Y B Suris
Technische Universität München
München, Germany

R J Szabo
Heriot-Watt University
Edinburgh, UK

S Tabachnikov
Pennsylvania State University
University Park, PA, USA

H Tasaki
Gakushuin University
Tokyo, Japan

M E Taylor
University of North Carolina
Chapel Hill, NC, USA

R Temam
Indiana University
Bloomington, IN, USA

B Temple
University of California at Davis
Davis, CA, USA

R P Thomas
Imperial College
London, UK

U Tillmann
University of Oxford
Oxford, UK

K P Tod
University of Oxford
Oxford, UK

J A Toth
McGill University
Montreal, QC, Canada

C A Tracy
University of California at Davis
Davis, CA, USA

A Trautman
Warsaw University
Warsaw, Poland

D Treschev
Moscow State University
Moscow, Russia

L Triolo
Università di Roma “Tor Vergata”
Rome, Italy

J Troost
Ecole Normale Supérieure
Paris, France

Tsou Sheung Tsun
University of Oxford
Oxford, UK

V Turaev
IRMA
Strasbourg, France

D Ueltschi
University of Arizona
Tucson, AZ, USA

A M Uranga
Consejo Superior de Investigaciones Científicas
Madrid, Spain

A Valentini
Perimeter Institute for Theoretical Physics
Waterloo, ON, Canada

M Vaugon
Université P.-M. Curie, Paris VI
Paris, France

P Di Vecchia
Nordita
Copenhagen, Denmark

A F Verbeure
Institute for Theoretical Physics
KU Leuven, Belgium

Y Colin de Verdière
Université de Grenoble 1
Saint-Martin d'Hères, France

M Viana
IMPA
Rio de Janeiro, Brazil

G Vitiello
Università degli Studi di Salerno
Baronissi (SA), Italy

D-V Voiculescu
University of California at Berkeley
Berkeley, CA, USA

S Waldmann
Albert-Ludwigs-Universität Freiburg
Freiburg, Germany

J Wambsganss
Universität Heidelberg
Heidelberg, Germany

R S Ward
University of Durham
Durham, UK

E Wayne
Boston University
Boston, MA, USA

F W Wehrli
University of Pennsylvania
Philadelphia, PA, USA

R F Werner
Technische Universität Braunschweig
Braunschweig, Germany

H Widom
University of California at Santa Cruz
Santa Cruz, CA, USA

C M Will
Washington University
St. Louis, MO, USA

N M J Woodhouse
University of Oxford
Oxford, UK

Siye Wu
University of Colorado
Boulder, CO, USA

V Wunsch
Friedrich-Schiller-Universität Jena
Jena, Germany

D R Yafaev
Université de Rennes
Rennes, France

M Yamada
Kyoto University
Kyoto, Japan

M Yuri
Hokkaido University
Sapporo, Japan

D Žubrinić
University of Zagreb
Zagreb, Croatia

V Županović
University of Zagreb
Zagreb, Croatia

R Zecchina
International Centre for Theoretical Physics (ICTP)
Trieste, Italy

S Zelditch
Johns Hopkins University
Baltimore, MD, USA

S Zelik
Universität Stuttgart
Stuttgart, Germany

S-C Zhang
Stanford University
Stanford, CA, USA

M B Ziane
University of Southern California
Los Angeles, CA, USA

M R Zirnbauer
Universität Köln
Köln, Germany

A Zumpano
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

CONTENTS LIST BY SUBJECT

Location references refer to the volume number and page number (separated by a colon).

INTRODUCTORY ARTICLES

- Classical Mechanics 1:1
- Differential Geometry 1:33
- Electromagnetism 1:40
- Equilibrium Statistical Mechanics 1:51
- Functional Analysis 1:88
- Minkowski Spacetime and Special Relativity 1:96
- Quantum Mechanics 1:109
- Topology 1:131

PHYSICS SUBJECTS

Classical Mechanics

- Boundary Control Method and Inverse Problems of Wave Propagation 1:340
- Constrained Systems 1:611
- Cotangent Bundle Reduction 1:658
- Gravitational N -body Problem (Classical) 2:575
- Hamiltonian Fluid Dynamics 2:593
- Hamiltonian Systems: Obstructions to Integrability 2:624
- Infinite-Dimensional Hamiltonian Systems 3:37
- Inverse Problem in Classical Mechanics 3:156
- KAM Theory and Celestial Mechanics 3:189
- Peakons 4:12
- Poisson Reduction 4:79
- Stability Problems in Celestial Mechanics 5:20
- Symmetry and Symplectic Reduction 5:190

Classical, Conformal and Topological Field Theory

- Topological Quantum Field Theory: Overview 5:278
- AdS/CFT Correspondence 1:174
- Axiomatic Approach to Topological Quantum Field Theory 1:232
- BF Theories 1:257
- Boundary Conformal Field Theory 1:333
- Chern–Simons Models: Rigorous Results 1:496

- Donaldson–Witten Theory 2:110
- Duality in Topological Quantum Field Theory 2:118
- Finite-Type Invariants 2:340
- Four-Manifold Invariants and Physics 2:386
- Gauge Theoretic Invariants of 4-Manifolds 2:457
- h -Pseudodifferential Operators and Applications 2:701
- The Jones Polynomial 3:179
- Knot Theory and Physics 3:220
- Kontsevich Integral 3:231
- Large- N and Topological Strings 3:263
- Mathai–Quillen Formalism 3:390
- Mathematical Knot Theory 3:399
- Operator Product Expansion in Quantum Field Theory 3:616
- Schwarz-Type Topological Quantum Field Theory 4:494
- Solitons and Other Extended Field Configurations 4:602
- Topological Defects and Their Homotopy Classification 5:257
- Topological Gravity, Two-Dimensional 5:264
- Topological Knot Theory and Macroscopic Physics 5:271
- Topological Sigma Models 5:290
- Two-Dimensional Conformal Field Theory and Vertex Operator Algebras 5:317
- WDVV Equations and Frobenius Manifolds 5:438

Condensed Matter and Optics

- Bose–Einstein Condensates 1:312
- Falicov–Kimball Model 2:283
- Fractional Quantum Hall Effect 2:402
- High T_c Superconductor Theory 2:645
- Hubbard Model 2:712
- Liquid Crystals 3:320
- Negative Refraction and Subdiffraction Imaging 3:483
- Nuclear Magnetic Resonance 3:592

Optical Caustics 3:620
 Quantum Phase Transitions 4:289
 Quasiperiodic Systems 4:308
 Renormalization: Statistical Mechanics and
 Condensed Matter 4:407
 Short-Range Spin Glasses: The Metastate
 Approach 4:570
 Topological Defects and Their Homotopy
 Classification 5:257

Disordered Systems

Cellular Automata 1:455
 Lagrangian Dispersion (Passive Scalar) 3:255
 Mean Field Spin Glasses and Neural
 Networks 3:407
 Percolation Theory 4:21
 Random Matrix Theory in Physics 4:338
 Random Walks in Random Environments 4:353
 Short-Range Spin Glasses: The Metastate
 Approach 4:570
 Spin Glasses 4:655
 Stochastic Loewner Evolutions 5:80
 Two-Dimensional Ising Model 5:322
 Wulff Droplets 5:462

Dynamical Systems

Averaging Methods 1:226
 Bifurcations of Periodic Orbits 1:285
 Billiards in Bounded Convex Domains 1:296
 Central Manifolds, Normal Forms 1:467
 Cellular Automata 1:455
 Chaos and Attractors 1:477
 Cotangent Bundle Reduction 1:658
 Diagrammatic Techniques in Perturbation
 Theory 2:54
 Dissipative Dynamical Systems of Infinite
 Dimension 2:101
 Dynamical Systems and Thermodynamics 2:125
 Dynamical Systems in Mathematical Physics:
 An Illustration from Water Waves 2:133
 Entropy and Quantitative Transversality 2:237
 Ergodic Theory 2:250
 Fractal Dimensions in Dynamics 2:394
 Generic Properties of Dynamical Systems 2:494
 Gravitational N -Body Problem (Classical) 2:575
 Hamiltonian Fluid Dynamics 2:593
 Hamiltonian Systems: Stability and Instability
 Theory 2:631
 Holomorphic Dynamics 2:652
 Homeomorphisms and Diffeomorphisms of the
 Circle 2:665
 Homoclinic Phenomena 2:672
 b -Pseudodifferential Operators and
 Applications 2:701
 Hyperbolic Billiards 2:716
 Hyperbolic Dynamical Systems 2:721
 Isomonodromic Deformations 3:173

KAM Theory and Celestial Mechanics 3:189
 Lyapunov Exponents and Strange Attractors 3:349
 Multiscale Approaches 3:465
 Normal Forms and Semiclassical
 Approximation 3:578
 Point-Vortex Dynamics 4:66
 Poisson Reduction 4:79
 Polygonal Billiards 4:84
 Quasiperiodic Systems 4:308
 Random Dynamical Systems 4:330
 Regularization For Dynamical ζ -Functions 4:386
 Resonances 4:415
 Riemann–Hilbert Problem 4:436
 Semiclassical Spectra and Closed Orbits 4:512
 Separatrix Splitting 4:535
 Stability Problems in Celestial Mechanics 5:20
 Stability Theory and KAM 5:26
 Symmetry and Symmetry Breaking in Dynamical
 Systems 5:184
 Symmetry and Symplectic Reduction 5:190
 Synchronization of Chaos 5:213
 Universality and Renormalization 5:343
 Weakly Coupled Oscillators 5:448

Equilibrium Statistical Mechanics

Bethe Ansatz 1:253
 Cluster Expansion 1:531
 Dimer Problems 2:61
 Eight Vertex and Hard Hexagon Models 2:155
 Falicov–Kimball Model 2:283
 Fermionic Systems 2:300
 Finitely Correlated States 2:334
 Holonomic Quantum Fields 2:660
 Hubbard Model 2:712
 Large Deviations in Equilibrium Statistical
 Mechanics 3:261
 Metastable States 3:417
 Phase Transitions in Continuous Systems 4:53
 Pirogov–Sinai Theory 4:60
 Quantum Central-Limit Theorems 4:130
 Quantum Phase Transitions 4:289
 Quantum Spin Systems 4:295
 Quantum Statistical Mechanics: Overview 4:302
 Reflection Positivity and Phase Transitions 4:376
 Short-Range Spin Glasses: The Metastate
 Approach 4:570
 Statistical Mechanics and Combinatorial
 Problems 5:50
 Statistical Mechanics of Interfaces 5:55
 Superfluids 5:115
 Toeplitz Determinants and Statistical
 Mechanics 5:244
 Two-Dimensional Ising Model 5:322
 Wulff Droplets 5:462

Fluid Dynamics

Bifurcations in Fluid Dynamics 1:281
 Breaking Water Waves 1:383

Capillary Surfaces 1:431
 Cauchy Problem for Burgers-Type Equations 1:446
 Compressible Flows: Mathematical Theory 1:595
 Fluid Mechanics: Numerical Methods 2:365
 Geophysical Dynamics 2:534
 Hamiltonian Fluid Dynamics 2:593
 Incompressible Euler Equations: Mathematical Theory 3:10
 Interfaces and Multicomponent Fluids 3:135
 Intermittency in Turbulence 3:144
 Inviscid Flows 3:160
 Korteweg–de Vries Equation and Other Modulation Equations 3:239
 Lagrangian Dispersion (Passive Scalar) 3:255
 Magnetohydrodynamics 3:375
 Newtonian Fluids and Thermohydraulics 3:492
 Non-Newtonian Fluids 3:560
 Partial Differential Equations: Some Examples 4:6
 Peakons 4:12
 Stability of Flows 5:1
 Superfluids 5:115
 Turbulence Theories 5:295
 Variational Methods in Turbulence 5:351
 Viscous Incompressible Fluids: Mathematical Theory 5:369
 Vortex Dynamics 5:390
 Wavelets: Application to Turbulence 5:408

Gauge Theory

Abelian and Nonabelian Gauge Theories Using Differential Forms 1:141
 Abelian Higgs Vortices 1:151
 AdS/CFT Correspondence 1:174
 Aharonov–Bohm Effect 1:191
 Anomalies 1:205
 BRST Quantization 1:386
 Chern–Simons Models: Rigorous Results 1:496
 Dirac Fields in Gravitation and Nonabelian Gauge Theory 2:67
 Donaldson–Witten Theory 2:110
 Effective Field Theories 2:139
 Electric–Magnetic Duality 2:201
 Electroweak Theory 2:209
 Exact Renormalization Group 2:272
 Gauge Theories from Strings 2:463
 Gauge Theory: Mathematical Applications 2:468
 Instantons: Topological Aspects 3:44
 Large- N and Topological Strings 3:263
 Lattice Gauge Theory 3:275
 Measure on Loop Spaces 3:413
 Noncommutative Geometry and the Standard Model 3:509
 Nonperturbative and Topological Aspects of Gauge Theory 3:568
 Perturbative Renormalization Theory and BRST 4:41
 Quantum Chromodynamics 4:144
 Quantum Electrodynamics and Its Precision Tests 4:168

Renormalization: General Theory 4:399
 Seiberg–Witten Theory 4:503
 Standard Model of Particle Physics 5:32
 Supergravity 5:122
 Supersymmetric Particle Models 5:140
 Symmetry Breaking in Field Theory 5:198
 Twistor Theory: Some Applications 5:303
 Two-Dimensional Models 5:328

General Relativity

General Relativity: Overview 2:487
 Asymptotic Structure and Conformal Infinity 1:221
 Black Hole Mechanics 1:300
 Boundaries for Spacetimes 1:326
 Brane Worlds 1:367
 Canonical General Relativity 1:412
 Critical Phenomena in Gravitational Collapse 1:668
 Computational Methods in General Relativity: The Theory 1:604
 Cosmology: Mathematical Aspects 1:653
 Dirac Fields in Gravitation and Nonabelian Gauge Theory 2:67
 Einstein–Cartan Theory 2:189
 Einstein’s Equations with Matter 2:195
 Einstein Equations: Exact Solutions 2:165
 Einstein Equations: Initial Value Formulation 2:173
 General Relativity: Experimental Tests 2:481
 Geometric Analysis and General Relativity 2:502
 Geometric Flows and the Penrose Inequality 2:510
 Gravitational Lensing 2:567
 Gravitational Waves 2:582
 Hamiltonian Reduction of Einstein’s Equations 2:607
 Minimal Submanifolds 3:420
 Newtonian Limit of General Relativity 3:503
 Quantum Field Theory in Curved Spacetime 4:202
 Relativistic Wave Equations Including Higher Spin Fields 4:391
 Shock Wave Refinement of the Friedman–Robertson–Walker Metric 4:559
 Spacetime Topology, Causal Structure and Singularities 4:617
 Spinors and Spin Coefficients 4:667
 Stability of Minkowski Space 5:14
 Stationary Black Holes 5:38
 Twistors 5:311

Integrable Systems

Integrable Systems: Overview 3:106
 Abelian Higgs Vortices 1:151
 Affine Quantum Groups 1:183
 Bäcklund Transformations 1:241

Bethe Ansatz 1:253
Bi-Hamiltonian Methods in Soliton Theory 1:290
Boundary-Value Problems For Integrable Equations 1:346
Calogero–Moser–Sutherland Systems of Nonrelativistic and Relativistic Type 1:403
 $\bar{\delta}$ -Approach to Integrable Systems 2:34
Eigenfunctions of Quantum Completely Integrable Systems 2:148
Functional Equations and Integrable Systems 2:425
Holonomic Quantum Fields 2:660
Instantons: Topological Aspects 3:44
Integrability and Quantum Field Theory 3:50
Integrable Discrete Systems 3:59
Integrable Systems and Algebraic Geometry 3:65
Integrable Systems and Discrete Geometry 3:78
Integrable Systems and Recursion Operators on Symplectic and Jacobi Manifolds 3:87
Integrable Systems and the Inverse Scattering Method 3:93
Integrable Systems in Random Matrix Theory 3:102
Isochronous Systems 3:166
Nonlinear Schrödinger Equations 3:552
Painlevé Equations 4:1
Peakons 4:12
Quantum Calogero–Moser Systems 4:123
Riemann–Hilbert Methods in Integrable Systems 4:429
Sine-Gordon Equation 4:576
Solitons and Kac–Moody Lie Algebras 4:594
Toda Lattices 5:235
Twistor Theory: Some Applications 5:303
Yang–Baxter Equations 5:465

M-Theory *see* String Theory and M-Theory

Nonequilibrium Statistical Mechanics

Nonequilibrium Statistical Mechanics (Stationary): Overview 3:530
Adiabatic Piston 1:160
Boltzmann Equation (Classical and Quantum) 1:306
Glassy Disordered Systems: Dynamical Evolution 2:553
Fourier Law 2:374
Interacting Particle Systems and Hydrodynamic Equations 3:123
Interacting Stochastic Particle Systems 3:130
Kinetic Equations 3:200
Macroscopic Fluctuations and Thermodynamic Functionals 3:357
Nonequilibrium Statistical Mechanics: Dynamical Systems Approach 3:540
Nonequilibrium Statistical Mechanics: Interaction between Theory and Numerical Simulations 3:544

Phase Transition Dynamics 4:47
Stochastic Resonance 5:86

Quantum Field Theory

Quantum Field Theory: A Brief Introduction 4:212
AdS/CFT Correspondence 1:174
Algebraic Approach to Quantum Field Theory 1:198
Anomalies 1:205
Axiomatic Quantum Field Theory 1:234
Batalin–Vilkovisky Quantization 1:247
Bosons and Fermions in External Fields 1:318
BRST Quantization 1:386
Constrained Systems 1:611
Constructive Quantum Field Theory 1:617
Current Algebra 1:674
Dirac Operator and Dirac Field 2:74
Dispersion Relations 2:87
Effective Field Theories 2:139
Electroweak Theory 2:209
Euclidean Field Theory 2:256
Exact Renormalization Group 2:272
Gerbes in Quantum Field Theory 2:539
Holonomic Quantum Fields 2:660
Hopf Algebra Structure of Renormalizable Quantum Field Theory 2:678
Indefinite Metric 3:17
Integrability and Quantum Field Theory 3:50
Large- N and Topological Strings 3:263
Nonperturbative and Topological Aspects of Gauge Theory 3:568
Operator Product Expansion in Quantum Field Theory 3:616
Quantum Fields with Indefinite Metric: Non-Trivial Models 4:216
Perturbation Theory and Its Techniques 4:28
Perturbative Renormalization Theory and BRST 4:41
Quantum Electrodynamics and Its Precision Tests 4:168
Quantum Fields with Topological Defects 4:221
Quantum Field Theory in Curved Spacetime 4:202
Quantum Phase Transitions 4:289
Renormalization: General Theory 4:399
Renormalization: Statistical Mechanics and Condensed Matter 4:407
Scattering, Asymptotic Completeness and Bound States 4:475
Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools 4:456
Scattering in Relativistic Quantum Field Theory: The Analytic Program 4:465
Seiberg–Witten Theory 4:503
Standard Model of Particle Physics 5:32
Supergravity 5:122
Supersymmetric Particle Models 5:140
Symmetries and Conservation Laws 5:166

- Symmetries in Quantum Field Theory: Algebraic Aspects 5:179
- Symmetries in Quantum Field Theory of Lower Spacetime Dimensions 5:172
- Symmetry Breaking in Field Theory 5:198
- Two-Dimensional Models 5:328
- Thermal Quantum Field Theory 5:227
- Tomita–Takesaki Modular Theory 5:251
- Topological Defects and Their Homotopy Classification 5:257
- Twistor Theory: Some Applications 5:303

Quantum Gravity

- Knot Invariants and Quantum Gravity 3:215
- Knot Theory and Physics 3:220
- Loop Quantum Gravity 3:339
- Quantum Cosmology 4:153
- Quantum Dynamics in Loop Quantum Gravity 4:165
- Quantum Field Theory in Curved Spacetime 4:202
- Quantum Geometry and Its Applications 4:230
- Spin Foams 4:645
- Wheeler–De Witt Theory 5:453

Quantum Information and Computation

- Capacities Enhanced By Entanglement 1:418
- Capacity for Quantum Information 1:424
- Channels in Quantum Information Theory 1:472
- Entanglement 2:228
- Entanglement Measures 2:233
- Finite Weyl Systems 2:328
- Optimal Cloning of Quantum States 3:628
- Quantum Channels: Classical Capacity 4:142
- Quantum Entropy 4:177
- Quantum Error Correction and Fault Tolerance 4:196
- Source Coding in Quantum Information Theory 4:609

Quantum Mechanics

- Aharonov–Bohm Effect 1:191
- Arithmetic Quantum Chaos 1:212
- Coherent States 1:537
- Geometric Phases 2:528
- \hbar -Pseudodifferential Operators and Applications 2:701
- N -particle Quantum Scattering 3:585
- Normal Forms and Semiclassical Approximation 3:578
- Quantum Entropy 4:177
- Quantum Ergodicity and Mixing of Eigenfunctions 4:183
- Quantum Mechanical Scattering Theory 4:251
- Quantum Mechanics: Foundations 4:260
- Quantum Mechanics: Generalizations 4:265

- Quantum Mechanics: Weak Measurements 4:276
- Quantum n -Body Problem 4:283
- Quantum Spin Systems 4:295
- Quasiperiodic Systems 4:308
- Schrödinger Operators 4:487
- Stability of Matter 5:8
- Stationary Phase Approximation 5:44
- Supersymmetric Quantum Mechanics 5:145
- Topological Defects and Their Homotopy Classification 5:257

String Theory and M -Theory

- AdS/CFT Correspondence 1:174
- Brane Construction of Gauge Theories 1:360
- Branes and Black Hole Statistical Mechanics 1:373
- Brane Worlds 1:367
- Calibrated Geometry and Special Lagrangian Submanifolds 1:398
- Compactification of Superstring Theory 1:586
- Derived Categories 2:41
- Fourier–Mukai Transform in String Theory 2:379
- Gauge Theories from Strings 2:463
- Large- N and Topological Strings 3:263
- Large- N Dualities 3:269
- Mirror Symmetry: A Geometric Survey 3:439
- Noncommutative Geometry from Strings 3:515
- Random Algebraic Geometry, Attractors and Flux Vacua 4:323
- Riemannian Holonomy Groups and Exceptional Holonomy 4:441
- String Field Theory 5:94
- String Theory: Phenomenology 5:103
- String Topology: Homotopy and Geometric Perspectives 5:111
- Superstring Theories 5:133
- Twistor Theory: Some Applications 5:303
- Two-Dimensional Conformal Field Theory and Vertex Operator Algebras 5:317

Yang–Mills Theory see Gauge Theory

RELATED MATHEMATICS SUBJECTS

Algebraic Techniques

- Affine Quantum Groups 1:183
- Braided and Modular Tensor Categories 1:351
- Clifford Algebras and Their Representations 1:518
- Derived Categories 2:41
- Finite-Dimensional Algebras and Quivers 2:313
- Finite Group Symmetry Breaking 2:322
- Hopf Algebras and Q -Deformation Quantum Groups 2:687
- Operads 3:609

Algebraic Topology

- Characteristic Classes 1:488
- Cohomology Theories 1:545
- Derived Categories 2:41
- Equivariant Cohomology and the Cartan Model 2:242
- Fourier–Mukai Transform in String Theory 2:379
- Index Theorems 3:23
- Intersection Theory 3:151
- K-theory 3:246
- Mathai–Quillen Formalism 3:390
- Operads 3:609
- Spectral Sequences 4:623
- String Topology: Homotopy and Geometric Perspectives 5:111

Complex Geometry

- Derived Categories 2:41
- Gauge Theory: Mathematical Applications 2:468
- Fourier–Mukai Transform in String Theory 2:379
- Knot Homologies 3:208
- Mirror Symmetry: A Geometric Survey 3:439
- Moduli Spaces: An Introduction 3:449
- Quillen Determinant 4:315
- Riemann Surfaces 4:419
- Riemann–Hilbert Problem 4:436
- Several Complex Variables: Basic Geometric Theory 4:540
- Several Complex Variables: Compact Manifolds 4:551
- Twistor Theory: Some Applications 5:303

Differential Geometry

- Calibrated Geometry and Special Lagrangian Submanifolds 1:398
- Capillary Surfaces 1:431
- Characteristic Classes 1:488
- Derived Categories 2:41
- Einstein Manifolds 2:182
- Fourier–Mukai Transform in String Theory 2:379
- Gauge Theory: Mathematical Applications 2:468
- Index Theorems 3:23
- Intersection Theory 3:151
- K-Theory 3:246
- Lorentzian Geometry 3:343
- Mathai–Quillen Formalism 3:390
- Moduli Spaces: An Introduction 3:449
- Quillen Determinant 4:315
- Pseudo-Riemannian Nilpotent Lie Groups 4:94
- Riemann–Hilbert Problem 4:436
- Riemannian Holonomy Groups and Exceptional Holonomy 4:441
- Singularity and Bifurcation Theory 4:588
- Supermanifolds 5:128
- Twistor Theory: Some Applications 5:303

Discrete Mathematics

- Arithmetic Quantum Chaos 1:212
- Combinatorics: Overview 1:553
- Number Theory in Physics 3:600
- Quasiperiodic Systems 4:308

Functional Analysis and Operator Algebras

- Bäcklund Transformations 1:241
- C*-Algebras and their Classification 1:393
- Coherent States 1:537
- Free Probability Theory 2:417
- Functional Integration in Quantum Physics 2:434
- Gauge Theory: Mathematical Applications 2:468
- \hbar -Pseudodifferential Operators and Applications 2:701
- The Jones Polynomial 3:179
- K-Theory 3:246
- Leray–Schauder Theory and Mapping Degree 3:281
- Ljusternik–Schnirelman Theory 3:328
- Ordinary Special Functions 3:637
- Positive Maps on C*-Algebras 4:88
- Quantum Dynamical Semigroups 4:159
- Saddle Point Problems 4:447
- Spectral Theory of Linear Operators 4:633
- Tomita–Takesaki Modular Theory 5:251
- von Neumann Algebras: Introduction, Modular Theory, and Classification Theory 5:379
- von Neumann Algebras: Subfactor Theory 5:385
- Wavelets: Applications 5:420
- Wavelets: Mathematical Theory 5:426

Lie Groups and Lie Algebras

- Classical Groups and Homogeneous Spaces 1:500
- Compact Groups and Their Representations 1:576
- Finite-Dimensional Algebras and Quivers 2:313
- Lie Groups: General Theory 3:286
- Lie Superalgebras and Their Representations 3:305
- Lie, Symplectic, and Poisson Groupoids and Their Lie Algebroids 3:312
- Pseudo-Riemannian Nilpotent Lie Groups 4:94
- Riemann–Hilbert Problem 4:436
- Solitons and Kac–Moody Lie Algebras 4:594

Low Dimensional Geometry

- Finite-type Invariants of 3-Manifolds 2:348
- Floer Homology 2:356
- Four-manifold Invariants and Physics 2:386
- Gauge Theoretic Invariants of 4-Manifolds 2:457
- Gauge Theory: Mathematical Applications 2:468
- The Jones Polynomial 3:179
- Knot Invariants and Quantum Gravity 3:215
- Large- N and Topological Strings 3:263

- Quantum 3-Manifold Invariants 4:117
- Singularities of the Ricci Flow 4:584
- Twistor Theory: Some Applications 5:303

Noncommutative Geometry

- Hopf Algebra Structure of Renormalizable Quantum Field Theory 2:678
- Noncommutative Geometry and the Standard Model 3:509
- Noncommutative Geometry from Strings 3:515
- Noncommutative Tori, Yang-Mills, and String Theory 3:524
- Path Integrals in Noncommutative Geometry 4:8
- Quantum Group Differentials, Bundles and Gauge Theory 4:236
- Quantum Hall Effect 4:244
- Riemann–Hilbert Problem 4:436

Ordinary and Partial Differential Equations

- Bifurcation Theory 1:275
- Boltzmann Equation (Classical and Quantum) 1:306
- Boundary Control Method and Inverse Problems of Wave Propagation 1:340
- Capillary Surfaces 1:431
- Cauchy Problem for Burgers-Type Equations 1:446
- Elliptic Differential Equations: Linear Theory 2:216
- Evolution Equations: Linear and Nonlinear 2:265
- Fluid Mechanics: Numerical Methods 2:365
- Ginzburg–Landau Equation 2:547
- Image Processing: Mathematics 3:1
- Inequalities in Sobolev Spaces 3:32
- Isomonodromic Deformations 3:173
- Kinetic Equations 3:200
- Localization For Quasiperiodic Potentials 3:333
- Magnetic Resonance Imaging 3:367
- Minimal Submanifolds 3:420
- Painlevé Equations 4:1
- Partial Differential Equations: Some Examples 4:6
- Relativistic Wave Equations Including Higher Spin Fields 4:391
- Riemann–Hilbert Problem 4:436
- Semilinear Wave Equations 4:518
- Separation of Variables for Differential Equations 4:526
- Stationary Phase Approximation 5:44
- Symmetric Hyperbolic Systems and Shock Waves 5:160
- Wave Equations and Diffraction 5:401

Quantization Methods and Path Integration

- Coherent States 1:537
- Deformation Quantization 2:1

- Deformation Quantization and Representation Theory 2:9
- Deformation Theory 2:16
- Deformations of the Poisson Bracket on a Symplectic Manifold 2:24
- Fedosov Quantization 2:291
- Feynman Path Integrals 2:307
- Functional Integration in Quantum Physics 2:434
- Path Integrals in Noncommutative Geometry 4:8
- Regularization for Dynamical ζ -Functions 4:386

Quantum Groups

- Affine Quantum Groups 1:183
- Bicrossproduct Hopf Algebras and Noncommutative Spacetime 1:265
- Braided and Modular Tensor Categories 1:351
- Classical r -Matrices, Lie Bialgebras, and Poisson Lie Groups 1:511
- Hopf Algebras and q -Deformation Quantum Groups 2:687
- Hopf Algebra Structure of Renormalizable Quantum Field Theory 2:678
- q -Special Functions 4:105
- Quantum Group Differentials, Bundles and Gauge Theory 4:236
- Yang–Baxter Equations 5:465

Stochastic Methods

- Determinantal Random Fields 2:47
- Free Probability Theory 2:417
- Growth Processes in Random Matrix Theory 2:586
- Integrable Systems in Random Matrix Theory 3:102
- Lagrangian Dispersion (Passive Scalar) 3:255
- Malliavin Calculus 3:383
- Measure on Loop Spaces 3:413
- Random Matrix Theory in Physics 4:338
- Random Partitions 4:347
- Random Walks in Random Environments 4:353
- Stochastic Differential Equations 5:63
- Stochastic Hydrodynamics 5:71
- Stochastic Loewner Evolutions 5:80
- Supersymmetry Methods in Random Matrix Theory 5:151
- Symmetry Classes in Random Matrix Theory 5:204

Symplectic Geometry and Topology

- Classical r -Matrices, Lie Bialgebras, and Poisson Lie Groups 1:511
- Contact Manifolds 1:631
- Deformations of the Poisson Bracket on a Symplectic Manifold 2:24
- Fedosov Quantization 2:291
- Floer Homology 2:356
- Graded Poisson Algebras 2:560

Hamiltonian Group Actions 2:600
Mirror Symmetry: A Geometric Survey 3:439
Multi-Hamiltonian Systems 3:459
Recursion Operators in Classical
Mechanics 4:371
Singularity and Bifurcation Theory 4:588
Stationary Phase Approximation 5:44

Variational Techniques

Capillary Surfaces 1:431
Control Problems in Mathematical Physics 1:636
Convex Analysis and Duality Methods 1:642

Free Interfaces and Free Discontinuities: Variational
Problems 2:411
 Γ -Convergence and Homogenization 2:449
Gauge Theory: Mathematical Applications 2:468
Geometric Measure Theory 2:520
Hamilton–Jacobi Equations and Dynamical
Systems: Variational Aspects 2:636
Minimax Principle in the Calculus of
Variations 3:432
Optimal Transportation 3:632
Variational Techniques for Ginzburg–Landau
Energies 5:355
Variational Techniques for Microstructures 5:363

CONTENTS

VOLUME 1

Introductory Article: Classical Mechanics	<i>G Gallavotti</i>	1
Introductory Article: Differential Geometry	<i>S Paycha</i>	33
Introductory Article: Electromagnetism	<i>N M J Woodhouse</i>	40
Introductory Article: Equilibrium Statistical Mechanics	<i>G Gallavotti</i>	51
Introductory Article: Functional Analysis	<i>S Paycha</i>	88
Introductory Article: Minkowski Spacetime and Special Relativity	<i>G L Naber</i>	96
Introductory Article: Quantum Mechanics	<i>G F dell'Antonio</i>	109
Introductory Article: Topology	<i>Tsou Sheung Tsun</i>	131

A

Abelian and Nonabelian Gauge Theories Using Differential Forms	<i>A C Hirshfeld</i>	141
Abelian Higgs Vortices	<i>J M Speight</i>	151
Adiabatic Piston	<i>Ch Gruber and A Lesne</i>	160
AdS/CFT Correspondence	<i>C P Herzog and I R Klebanov</i>	174
Affine Quantum Groups	<i>G W Delius and N MacKay</i>	183
Aharonov–Bohm Effect	<i>M Socolovsky</i>	191
Algebraic Approach to Quantum Field Theory	<i>R Brunetti and K Fredenhagen</i>	198
Anderson Localization	<i>see Localization for Quasiperiodic Potentials</i>	
Anomalies	<i>S L Adler</i>	205
Arithmetic Quantum Chaos	<i>J Marklof</i>	212
Asymptotic Structure and Conformal Infinity	<i>J Frauendiener</i>	221
Averaging Methods	<i>A I Neishtadt</i>	226
Axiomatic Approach to Topological Quantum Field Theory	<i>C Blanchet and V Turaev</i>	232
Axiomatic Quantum Field Theory	<i>B Kuckert</i>	234

B

Bäcklund Transformations	<i>D Levi</i>	241
Batalin–Vilkovisky Quantization	<i>A C Hirshfeld</i>	247
Bethe Ansatz	<i>M T Batchelor</i>	253
BF Theories	<i>M Blau</i>	257
Bicrossproduct Hopf Algebras and Noncommutative Spacetime	<i>S Majid</i>	265

Bifurcation Theory	<i>M Haragus and G Iooss</i>	275
Bifurcations in Fluid Dynamics	<i>G Schneider</i>	281
Bifurcations of Periodic Orbits	<i>J-P Francoise</i>	285
Bi-Hamiltonian Methods in Soliton Theory	<i>M Pedroni</i>	290
Billiards in Bounded Convex Domains	<i>S Tabachnikov</i>	296
Black Hole Mechanics	<i>A Ashtekar</i>	300
Boltzmann Equation (Classical and Quantum)	<i>M Pulvirenti</i>	306
Bose–Einstein Condensates	<i>F Dalfovo, L P Pitaevskii and S Stringari</i>	312
Bosons and Fermions in External Fields	<i>E Langmann</i>	318
Boundaries for Spacetimes	<i>S G Harris</i>	326
Boundary Conformal Field Theory	<i>J Cardy</i>	333
Boundary Control Method and Inverse Problems of Wave Propagation	<i>M I Belishev</i>	340
Boundary-Value Problems for Integrable Equations	<i>B Pelloni</i>	346
Braided and Modular Tensor Categories	<i>V Lyubashenko</i>	351
Brane Construction of Gauge Theories	<i>S L Cacciatori</i>	360
Brane Worlds	<i>R Maartens</i>	367
Branes and Black Hole Statistical Mechanics	<i>S R Das</i>	373
Breaking Water Waves	<i>A Constantin</i>	383
BRST Quantization	<i>M Henneaux</i>	386

C

C*-Algebras and their Classification	<i>G A Elliott</i>	393
Calibrated Geometry and Special Lagrangian Submanifolds	<i>D D Joyce</i>	398
Calogero–Moser–Sutherland Systems of Nonrelativistic and Relativistic Type	<i>S N M Ruijsenaars</i>	403
Canonical General Relativity	<i>C Rovelli</i>	412
Capacities Enhanced by Entanglement	<i>P Hayden</i>	418
Capacity for Quantum Information	<i>D Kretschmann</i>	424
Capillary Surfaces	<i>R Finn</i>	431
Cartan Model	<i>see Equivariant Cohomology and the Cartan Model</i>	
Cauchy Problem for Burgers-Type Equations	<i>G M Henkin</i>	446
Cellular Automata	<i>M Bruschi and F Musso</i>	455
Central Manifolds, Normal Forms	<i>P Bonckaert</i>	467
Channels in Quantum Information Theory	<i>M Keyl</i>	472
Chaos and Attractors	<i>R Gilmore</i>	477
Characteristic Classes	<i>P B Gilkey, R Ivanova and S Nikčević</i>	488
Chern–Simons Models: Rigorous Results	<i>A N Sengupta</i>	496
Classical Groups and Homogeneous Spaces	<i>S Gindikin</i>	500
Classical <i>r</i> -Matrices, Lie Bialgebras, and Poisson Lie Groups	<i>M A Semenov-Tian-Shansky</i>	511
Clifford Algebras and Their Representations	<i>A Trautman</i>	518
Cluster Expansion	<i>R Kotecký</i>	531
Coherent States	<i>S T Ali</i>	537
Cohomology Theories	<i>U Tillmann</i>	545
Combinatorics: Overview	<i>C Krattenthaler</i>	553
Compact Groups and Their Representations	<i>A Kirillov and A Kirillov, Jr.</i>	576
Compactification of Superstring Theory	<i>M R Douglas</i>	586
Compressible Flows: Mathematical Theory	<i>G-Q Chen</i>	595
Computational Methods in General Relativity: The Theory	<i>M W Choptuik</i>	604

Confinement	<i>see</i> Quantum Chromodynamics	
Conformal Geometry	<i>see</i> Two-dimensional Conformal Field Theory and Vertex Operator Algebras	
Conservation Laws	<i>see</i> Symmetries and Conservation Laws	
Constrained Systems	<i>M Henneaux</i>	611
Constructive Quantum Field Theory	<i>G Gallavotti</i>	617
Contact Manifolds	<i>J B Etnyre</i>	631
Control Problems in Mathematical Physics	<i>B Piccoli</i>	636
Convex Analysis and Duality Methods	<i>G Bouchitté</i>	642
Cosmic Censorship	<i>see</i> Spacetime Topology, Causal Structure and Singularities	
Cosmology: Mathematical Aspects	<i>G F R Ellis</i>	653
Cotangent Bundle Reduction	<i>J-P Ortega and T S Ratiu</i>	658
Critical Phenomena in Gravitational Collapse	<i>C Gundlach</i>	668
Current Algebra	<i>G A Goldin</i>	674

VOLUME 2

D		
Deformation Quantization	<i>A C Hirshfeld</i>	1
Deformation Quantization and Representation Theory	<i>S Waldmann</i>	9
Deformation Theory	<i>M J Pflaum</i>	16
Deformations of the Poisson Bracket on a Symplectic Manifold	<i>S Gutt and S Waldmann</i>	24
$\bar{\partial}$ -Approach to Integrable Systems	<i>P G Grinevich</i>	34
Derived Categories	<i>E R Sharpe</i>	41
Determinantal Random Fields	<i>A Soshnikov</i>	47
Diagrammatic Techniques in Perturbation Theory	<i>G Gentile</i>	54
Dimer Problems	<i>R Kenyon</i>	61
Dirac Fields in Gravitation and Nonabelian Gauge Theory	<i>J A Smoller</i>	67
Dirac Operator and Dirac Field	<i>S N M Ruijsenaars</i>	74
Dispersion Relations	<i>J Bros</i>	87
Dissipative Dynamical Systems of Infinite Dimension	<i>M Efendiev, S Zelik and A Miranville</i>	101
Donaldson Invariants	<i>see</i> Gauge Theoretic Invariants of 4-Manifolds	
Donaldson–Witten Theory	<i>M Mariño</i>	110
Duality in Topological Quantum Field Theory	<i>C Lozano and J M F Labastida</i>	118
Dynamical Systems and Thermodynamics	<i>A Carati, L Galgani and A Giorgilli</i>	125
Dynamical Systems in Mathematical Physics: An Illustration from Water Waves	<i>O Goubet</i>	133
E		
Effective Field Theories	<i>G Ecker</i>	139
Eigenfunctions of Quantum Completely Integrable Systems	<i>J A Toth</i>	148
Eight Vertex and Hard Hexagon Models	<i>P A Pearce</i>	155
Einstein Equations: Exact Solutions	<i>Jiří Bičák</i>	165
Einstein Equations: Initial Value Formulation	<i>J Isenberg</i>	173
Einstein Manifolds	<i>A S Dancer</i>	182
Einstein–Cartan Theory	<i>A Trautman</i>	189
Einstein’s Equations with Matter	<i>Y Choquet-Bruhat</i>	195
Electric–Magnetic Duality	<i>Tsou Sheung Tsun</i>	201
Electroweak Theory	<i>K Konishi</i>	209
Elliptic Differential Equations: Linear Theory	<i>C Amrouche, M Krbec, Š Nečasová and B Lucquin-Desreux</i>	216
Entanglement	<i>R F Werner</i>	228

Entanglement Measures	<i>R F Werner</i>	233
Entropy and Quantitative Transversality	<i>G Comte</i>	237
Equivariant Cohomology and the Cartan Model	<i>E Meinrenken</i>	242
Ergodic Theory	<i>M Yuri</i>	250
Euclidean Field Theory	<i>F Guerra</i>	256
Evolution Equations: Linear and Nonlinear	<i>J Escher</i>	265
Exact Renormalization Group	<i>P K Mitter</i>	272

F		
Falicov–Kimball Model	<i>Ch Gruber and D Ueltschi</i>	283
Fedosov Quantization	<i>N Neumaier</i>	291
Feigenbaum Phenomenon	<i>see Universality and Renormalization</i>	
Fermionic Systems	<i>V Mastropietro</i>	300
Feynman Path Integrals	<i>S Mazzucchi</i>	307
Finite-Dimensional Algebras and Quivers	<i>A Savage</i>	313
Finite Group Symmetry Breaking	<i>G Gaeta</i>	322
Finite Weyl Systems	<i>D-M Schlingemann</i>	328
Finitely Correlated States	<i>R F Werner</i>	334
Finite-Type Invariants	<i>D Bar-Natan</i>	340
Finite-Type Invariants of 3-Manifolds	<i>T T Q Lê</i>	348
Floer Homology	<i>P B Kronheimer</i>	356
Fluid Mechanics: Numerical Methods	<i>J-L Guermond</i>	365
Fourier Law	<i>F Bonetto and L Rey-Bellet</i>	374
Fourier–Mukai Transform in String Theory	<i>B Andreas</i>	379
Four-Manifold Invariants and Physics	<i>C Nash</i>	386
Fractal Dimensions in Dynamics	<i>V Županović and D Žubrinić</i>	394
Fractional Quantum Hall Effect	<i>J K Jain</i>	402
Free Interfaces and Free Discontinuities: Variational Problems	<i>G Buttazzo</i>	411
Free Probability Theory	<i>D-V Voiculescu</i>	417
Frobenius Manifolds	<i>see WDVV Equations and Frobenius Manifolds</i>	
Functional Equations and Integrable Systems	<i>H W Braden</i>	425
Functional Integration in Quantum Physics	<i>C DeWitt-Morette</i>	434

G		
Γ -Convergence and Homogenization	<i>G Dal Maso</i>	449
Gauge Theoretic Invariants of 4-Manifolds	<i>S Bauer</i>	457
Gauge Theories from Strings	<i>P Di Vecchia</i>	463
Gauge Theory: Mathematical Applications	<i>S K Donaldson</i>	468
General Relativity: Experimental Tests	<i>C M Will</i>	481
General Relativity: Overview	<i>R Penrose</i>	487
Generic Properties of Dynamical Systems	<i>C Bonatti</i>	494
Geometric Analysis and General Relativity	<i>L Andersson</i>	502
Geometric Flows and the Penrose Inequality	<i>H Bray</i>	510
Geometric Measure Theory	<i>G Albierti</i>	520
Geometric Phases	<i>P Lévy</i>	528
Geophysical Dynamics	<i>M B Ziane</i>	534
Gerbes in Quantum Field Theory	<i>J Mickelsson</i>	539

Ginzburg–Landau Equation	<i>Y Morita</i>	547
Glassy Disordered Systems: Dynamical Evolution	<i>S Franz</i>	553
Graded Poisson Algebras	<i>A S Cattaneo, D Fiorenza and R Longoni</i>	560
Gravitational Lensing	<i>J Wambsganss</i>	567
Gravitational N -Body Problem (Classical)	<i>D C Heggie</i>	575
Gravitational Waves	<i>G González and J Pullin</i>	582
Growth Processes in Random Matrix Theory	<i>K Johansson</i>	586

H		
Hamiltonian Fluid Dynamics	<i>P J Morrison</i>	593
Hamiltonian Group Actions	<i>L C Jeffrey</i>	600
Hamiltonian Reduction of Einstein’s Equations	<i>A E Fischer and V Moncrief</i>	607
Hamiltonian Systems: Obstructions to Integrability	<i>M Irigoyen</i>	624
Hamiltonian Systems: Stability and Instability Theory	<i>P Bernard</i>	631
Hamilton–Jacobi Equations and Dynamical Systems: Variational Aspects	<i>A Siconolfi</i>	636
Hard Hexagon Model	<i>see Eight Vertex and Hard Hexagon Models</i>	
High T_c Superconductor Theory	<i>S-C Zhang</i>	645
Holomorphic Dynamics	<i>M Lyubich</i>	652
Holonomic Quantum Fields	<i>J Palmer</i>	660
Homeomorphisms and Diffeomorphisms of the Circle	<i>A Zumpano and A Sarmiento</i>	665
Homoclinic Phenomena	<i>S E Newhouse</i>	672
Hopf Algebra Structure of Renormalizable Quantum Field Theory	<i>D Kreimer</i>	678
Hopf Algebras and q -Deformation Quantum Groups	<i>S Majid</i>	687
\hbar -Pseudodifferential Operators and Applications	<i>B Helffer</i>	701
Hubbard Model	<i>H Tasaki</i>	712
Hydrodynamic Equations	<i>see Interacting Particle Systems and Hydrodynamic Equations</i>	
Hyperbolic Billiards	<i>M P Wojtkowski</i>	716
Hyperbolic Dynamical Systems	<i>B Hasselblatt</i>	721

VOLUME 3

I		
Image Processing: Mathematics	<i>G Aubert and P Kornprobst</i>	1
Incompressible Euler Equations: Mathematical Theory	<i>D Chae</i>	10
Indefinite Metric	<i>H Gottschalk</i>	17
Index Theorems	<i>P B Gilkey, K Kirsten, R Ivanova and J H Park</i>	23
Inequalities in Sobolev Spaces	<i>M Vaugon</i>	32
Infinite-Dimensional Hamiltonian Systems	<i>R Schmid</i>	37
Instantons: Topological Aspects	<i>M Jardim</i>	44
Integrability and Quantum Field Theory	<i>T J Hollowood</i>	50
Integrable Discrete Systems	<i>O Ragnisco</i>	59
Integrable Systems and Algebraic Geometry	<i>E Previato</i>	65
Integrable Systems and Discrete Geometry	<i>A Doliwa and P M Santini</i>	78
Integrable Systems and Recursion Operators on Symplectic and Jacobi Manifolds	<i>R Caseiro and J M Nunes da Costa</i>	87
Integrable Systems and the Inverse Scattering Method	<i>A S Fokas</i>	93
Integrable Systems in Random Matrix Theory	<i>C A Tracy and H Widom</i>	102
Integrable Systems: Overview	<i>Francesco Calogero</i>	106

Interacting Particle Systems and Hydrodynamic Equations	<i>C Landim</i>	123
Interacting Stochastic Particle Systems	<i>H Spohn</i>	130
Interfaces and Multicomponent Fluids	<i>J Kim and J Lowengrub</i>	135
Intermittency in Turbulence	<i>J Jiménez</i>	144
Intersection Theory	<i>A Kresch</i>	151
Inverse Problem in Classical Mechanics	<i>R G Novikov</i>	156
Inverse Problems in Wave Propagation	<i>see Boundary Control Method and Inverse Problems of Wave Propagation</i>	
Inviscid Flows	<i>R Robert</i>	160
Ising Model	<i>see Two-Dimensional Ising Model</i>	
Isochronous Systems	<i>Francesco Calogero</i>	166
Isomonodromic Deformations	<i>V P Kostov</i>	173
J		
The Jones Polynomial	<i>V F R Jones</i>	179
K		
Kac–Moody Lie Algebras	<i>see Solitons and Kac–Moody Lie Algebras</i>	
KAM Theory and Celestial Mechanics	<i>L Chierchia</i>	189
Kinetic Equations	<i>C Bardos</i>	200
Knot Homologies	<i>J Rasmussen</i>	208
Knot Invariants and Quantum Gravity	<i>R Gambini and J Pullin</i>	215
Knot Theory and Physics	<i>L H Kauffman</i>	220
Kontsevich Integral	<i>S Chmutov and S Duzhin</i>	231
Korteweg–de Vries Equation and Other Modulation Equations	<i>G Schneider and E Wayne</i>	239
K-Theory	<i>V Mathai</i>	246
L		
Lagrangian Dispersion (Passive Scalar)	<i>G Falkovich</i>	255
Large Deviations in Equilibrium Statistical Mechanics	<i>S Shlosman</i>	261
Large- <i>N</i> and Topological Strings	<i>R Gopakumar</i>	263
Large- <i>N</i> Dualities	<i>A Grassi</i>	269
Lattice Gauge Theory	<i>A Di Giacomo</i>	275
Leray–Schauder Theory and Mapping Degree	<i>J Mawhin</i>	281
Lie Bialgebras	<i>see Classical <i>r</i>-Matrices, Lie Bialgebras, and Poisson Lie Groups</i>	
Lie Groups: General Theory	<i>R Gilmore</i>	286
Lie Superalgebras and Their Representations	<i>L Frappat</i>	305
Lie, Symplectic, and Poisson Groupoids and Their Lie Algebroids	<i>C-M Marle</i>	312
Liquid Crystals	<i>O D Lavrentovich</i>	320
Ljusternik–Schnirelman Theory	<i>J Mawhin</i>	328
Localization for Quasiperiodic Potentials	<i>S Jitomirskaya</i>	333
Loop Quantum Gravity	<i>C Rovelli</i>	339
Lorentzian Geometry	<i>P E Ehrlich and S B Kim</i>	343
Lyapunov Exponents and Strange Attractors	<i>M Viana</i>	349
M		
Macroscopic Fluctuations and Thermodynamic Functionals	<i>G Jona-Lasinio</i>	357
Magnetic Resonance Imaging	<i>C L Epstein and F W Wehrli</i>	367
Magnetohydrodynamics	<i>C Le Bris</i>	375

Malliavin Calculus	<i>A B Cruzeiro</i>	383
Marsden–Weinstein Reduction	<i>see</i> Cotangent Bundle Reduction: Poisson Reduction: Symmetry and Symplectic Reduction	
Maslov Index	<i>see</i> Optical Caustics: Semiclassical Spectra and Closed Orbits: Stationary Phase Approximation	
Mathai–Quillen Formalism	<i>S Wu</i>	390
Mathematical Knot Theory	<i>L Boi</i>	399
Matrix Product States	<i>see</i> Finitely Correlated States	
Mean Curvature Flow	<i>see</i> Geometric Flows and the Penrose Inequality	
Mean Field Spin Glasses and Neural Networks	<i>A Bovier</i>	407
Measure on Loop Spaces	<i>H Airault</i>	413
Metastable States	<i>S Shlosman</i>	417
Minimal Submanifolds	<i>T H Colding and W P Minicozzi II</i>	420
Minimax Principle in the Calculus of Variations	<i>A Abbondandolo</i>	432
Mirror Symmetry: A Geometric Survey	<i>R P Thomas</i>	439
Modular Tensor Categories	<i>see</i> Braided and Modular Tensor Categories	
Moduli Spaces: An Introduction	<i>F Kirwan</i>	449
Multicomponent Fluids	<i>see</i> Interfaces and Multicomponent Fluids	
Multi-Hamiltonian Systems	<i>F Magri and M Pedroni</i>	459
Multiscale Approaches	<i>A Lesne</i>	465

N

Negative Refraction and Subdiffraction Imaging	<i>S O'Brien and S A Ramakrishna</i>	483
Newtonian Fluids and Thermohydraulics	<i>G Labrosse and G Kasperski</i>	492
Newtonian Limit of General Relativity	<i>J Ehlers</i>	503
Noncommutative Geometry and the Standard Model	<i>T Schücker</i>	509
Noncommutative Geometry from Strings	<i>Chong-Sun Chu</i>	515
Noncommutative Tori, Yang–Mills, and String Theory	<i>A Konechny</i>	524
Nonequilibrium Statistical Mechanics (Stationary): Overview	<i>G Gallavotti</i>	530
Nonequilibrium Statistical Mechanics: Dynamical Systems Approach	<i>P Buttà and C Marchioro</i>	540
Nonequilibrium Statistical Mechanics: Interaction between Theory and Numerical Simulations	<i>R Livi</i>	544
Nonlinear Schrödinger Equations	<i>M J Ablowitz and B Prinari</i>	552
Non-Newtonian Fluids	<i>C Guillopé</i>	560
Nonperturbative and Topological Aspects of Gauge Theory	<i>R W Jackiw</i>	568
Normal Forms and Semiclassical Approximation	<i>D Bambusi</i>	578
N-Particle Quantum Scattering	<i>D R Yafaev</i>	585
Nuclear Magnetic Resonance	<i>P T Callaghan</i>	592
Number Theory in Physics	<i>M Marcolli</i>	600

O

Operads	<i>J Stasheff</i>	609
Operator Product Expansion in Quantum Field Theory	<i>H Osborn</i>	616
Optical Caustics	<i>A Joets</i>	620
Optimal Cloning of Quantum States	<i>M Keyl</i>	628
Optimal Transportation	<i>Y Brenier</i>	632
Ordinary Special Functions	<i>W Van Assche</i>	637

VOLUME 4

P

Painlevé Equations	<i>N Joshi</i>	1
Partial Differential Equations: Some Examples	<i>R Temam</i>	6
Path Integral Methods	<i>see Functional Integration in Quantum Physics; Feynman Path Integrals</i>	
Path Integrals in Noncommutative Geometry	<i>R Léandre</i>	8
Peakons	<i>D D Holm</i>	12
Penrose Inequality	<i>see Geometric Flows and the Penrose Inequality</i>	
Percolation Theory	<i>V Beffara and V Sidoravicius</i>	21
Perturbation Theory and Its Techniques	<i>R J Szabo</i>	28
Perturbative Renormalization Theory and BRST	<i>K Fredenhagen and M Dütsch</i>	41
Phase Transition Dynamics	<i>A Onuki</i>	47
Phase Transitions in Continuous Systems	<i>E Presutti</i>	53
Pirogov–Sinai Theory	<i>R Kotecký</i>	60
Point-Vortex Dynamics	<i>S Boatto and D Crowdy</i>	66
Poisson Lie Groups	<i>see Classical r-Matrices, Lie Bialgebras, and Poisson Lie Groups</i>	
Poisson Reduction	<i>J-P Ortega and T S Ratiu</i>	79
Polygonal Billiards	<i>S Tabachnikov</i>	84
Positive Maps on C -Algebras	<i>F Cipriani</i>	88
Pseudo-Riemannian Nilpotent Lie Groups	<i>P E Parker</i>	94

Q

q -Special Functions	<i>T H Koornwinder</i>	105
Quantum 3-Manifold Invariants	<i>C Blanchet and V Turaev</i>	117
Quantum Calogero–Moser Systems	<i>R Sasaki</i>	123
Quantum Central-Limit Theorems	<i>A F Verbeure</i>	130
Quantum Channels: Classical Capacity	<i>A S Holevo</i>	142
Quantum Chromodynamics	<i>G Sterman</i>	144
Quantum Cosmology	<i>M Bojowald</i>	153
Quantum Dynamical Semigroups	<i>R Alicki</i>	159
Quantum Dynamics in Loop Quantum Gravity	<i>H Sahlmann</i>	165
Quantum Electrodynamics and Its Precision Tests	<i>S Laporta and E Remiddi</i>	168
Quantum Entropy	<i>D Petz</i>	177
Quantum Ergodicity and Mixing of Eigenfunctions	<i>S Zelditch</i>	183
Quantum Error Correction and Fault Tolerance	<i>D Gottesman</i>	196
Quantum Field Theory in Curved Spacetime	<i>B S Kay</i>	202
Quantum Field Theory: A Brief Introduction	<i>L H Ryder</i>	212
Quantum Fields with Indefinite Metric: Non-Trivial Models	<i>S Albeverio and H Gottschalk</i>	216
Quantum Fields with Topological Defects	<i>M Blasone, G Vitiello and P Jizba</i>	221
Quantum Geometry and Its Applications	<i>A Ashtekar and J Lewandowski</i>	230
Quantum Group Differentials, Bundles and Gauge Theory	<i>T Brzeziński</i>	236
Quantum Hall Effect	<i>K Hannabuss</i>	244
Quantum Mechanical Scattering Theory	<i>D R Yafaev</i>	251
Quantum Mechanics: Foundations	<i>R Penrose</i>	260
Quantum Mechanics: Generalizations	<i>P Pearle and A Valentini</i>	265
Quantum Mechanics: Weak Measurements	<i>L Diósi</i>	276
Quantum n -Body Problem	<i>R G Littlejohn</i>	283

Quantum Phase Transitions	<i>S Sachdev</i>	289
Quantum Spin Systems	<i>B Nachtergaele</i>	295
Quantum Statistical Mechanics: Overview	<i>L Triolo</i>	302
Quasiperiodic Systems	<i>P Kramer</i>	308
Quillen Determinant	<i>S Scott</i>	315
Quivers	see Finite-Dimensional Algebras and Quivers	
R		
Random Algebraic Geometry, Attractors and Flux Vacua	<i>M R Douglas</i>	323
Random Dynamical Systems	<i>V Araújo</i>	330
Random Matrix Theory in Physics	<i>T Guhr</i>	338
Random Partitions	<i>A Okounkov</i>	347
Random Walks in Random Environments	<i>L V Bogachev</i>	353
Recursion Operators in Classical Mechanics	<i>F Magri and M Pedroni</i>	371
Reflection Positivity and Phase Transitions	<i>Y Kondratiev and Y Kozitsky</i>	376
Regularization for Dynamical ζ -Functions	<i>V Baladi</i>	386
Relativistic Wave Equations Including Higher Spin Fields	<i>R Illge and V Wunsch</i>	391
Renormalization: General Theory	<i>J C Collins</i>	399
Renormalization: Statistical Mechanics and Condensed Matter	<i>M Salmhofer</i>	407
Resonances	<i>N Burq</i>	415
Ricci Flow	see Singularities of the Ricci Flow	
Riemann Surfaces	<i>K Hulek</i>	419
Riemann–Hilbert Methods in Integrable Systems	<i>D Shepelsky</i>	429
Riemann–Hilbert Problem	<i>V P Kostov</i>	436
Riemannian Holonomy Groups and Exceptional Holonomy	<i>D D Joyce</i>	441
S		
Saddle Point Problems	<i>M Schechter</i>	447
Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools	<i>D Buchholz and S J Summers</i>	456
Scattering in Relativistic Quantum Field Theory: The Analytic Program	<i>J Bros</i>	465
Scattering, Asymptotic Completeness and Bound States	<i>D Iagolnitzer and J Magnen</i>	475
Schrödinger Operators	<i>V Bach</i>	487
Schwarz-Type Topological Quantum Field Theory	<i>R K Kaul, T R Govindarajan and P Ramadevi</i>	494
Seiberg–Witten Theory	<i>Siye Wu</i>	503
Semiclassical Approximation	see Stationary Phase Approximation; Normal Forms and Semiclassical Approximation	
Semiclassical Spectra and Closed Orbits	<i>Y Colin de Verdière</i>	512
Semilinear Wave Equations	<i>P D’Ancona</i>	518
Separation of Variables for Differential Equations	<i>S Rauch-Wojciechowski and K Marciniak</i>	526
Separatrix Splitting	<i>D Treschev</i>	535
Several Complex Variables: Basic Geometric Theory	<i>A Huckleberry and T Peternell</i>	540
Several Complex Variables: Compact Manifolds	<i>A Huckleberry and T Peternell</i>	551
Shock Wave Refinement of the Friedman–Robertson–Walker Metric	<i>B Temple and J Smoller</i>	559
Shock Waves	see Symmetric Hyperbolic Systems and Shock Waves	
Short-Range Spin Glasses: The Metastate Approach	<i>C M Newman and D L Stein</i>	570
Sine-Gordon Equation	<i>S N M Ruijsenaars</i>	576
Singularities of the Ricci Flow	<i>M Anderson</i>	584
Singularity and Bifurcation Theory	<i>J-P Françoise and C Piquet</i>	588

Sobolev Spaces <i>see</i> Inequalities in Sobolev Spaces	
Solitons and Kac–Moody Lie Algebras <i>E Date</i>	594
Solitons and Other Extended Field Configurations <i>R S Ward</i>	602
Source Coding in Quantum Information Theory <i>N Datta and T C Dorlas</i>	609
Spacetime Topology, Causal Structure and Singularities <i>R Penrose</i>	617
Special Lagrangian Submanifolds <i>see</i> Calibrated Geometry and Special Lagrangian Submanifolds	
Spectral Sequences <i>P Selick</i>	623
Spectral Theory of Linear Operators <i>M Schechter</i>	633
Spin Foams <i>A Perez</i>	645
Spin Glasses <i>F Guerra</i>	655
Spinors and Spin Coefficients <i>K P Tod</i>	667

VOLUME 5

Stability of Flows <i>S Friedlander</i>	1
Stability of Matter <i>J P Solovej</i>	8
Stability of Minkowski Space <i>S Klainerman</i>	14
Stability Problems in Celestial Mechanics <i>A Celletti</i>	20
Stability Theory and KAM <i>G Gentile</i>	26
Standard Model of Particle Physics <i>G Altarelli</i>	32
Stationary Black Holes <i>R Beig and P T Chruściel</i>	38
Stationary Phase Approximation <i>J J Duistermaat</i>	44
Statistical Mechanics and Combinatorial Problems <i>R Zecchina</i>	50
Statistical Mechanics of Interfaces <i>S Miracle-Solé</i>	55
Stochastic Differential Equations <i>F Russo</i>	63
Stochastic Hydrodynamics <i>B Ferrario</i>	71
Stochastic Loewner Evolutions <i>G F Lawler</i>	80
Stochastic Resonance <i>S Herrmann and P Imkeller</i>	86
Strange Attractors <i>see</i> Lyapunov Exponents and Strange Attractors	
String Field Theory <i>L Rastelli</i>	94
String Theory: Phenomenology <i>A M Uranga</i>	103
String Topology: Homotopy and Geometric Perspectives <i>R L Cohen</i>	111
Superfluids <i>D Einzel</i>	115
Supergravity <i>K S Stelle</i>	122
Supermanifolds <i>F A Rogers</i>	128
Superstring Theories <i>C Bachas and J Troost</i>	133
Supersymmetric Particle Models <i>S Pokorski</i>	140
Supersymmetric Quantum Mechanics <i>J-W van Holten</i>	145
Supersymmetry Methods in Random Matrix Theory <i>M R Zirnbauer</i>	151
Symmetric Hyperbolic Systems and Shock Waves <i>S Kichenassamy</i>	160
Symmetries and Conservation Laws <i>L H Ryder</i>	166
Symmetries in Quantum Field Theory of Lower Spacetime Dimensions <i>J Mund and K-H Rehren</i>	172
Symmetries in Quantum Field Theory: Algebraic Aspects <i>J E Roberts</i>	179
Symmetry and Symmetry Breaking in Dynamical Systems <i>I Melbourne</i>	184
Symmetry and Symplectic Reduction <i>J-P Ortega and T S Ratiu</i>	190
Symmetry Breaking in Field Theory <i>T W B Kibble</i>	198
Symmetry Classes in Random Matrix Theory <i>M R Zirnbauer</i>	204
Synchronization of Chaos <i>M A Aziz-Alaoui</i>	213

T

t Hooft–Polyakov Monopoles *see* Solitons and Other Extended Field Configurations

Thermal Quantum Field Theory *C D Jäkel* 227

Thermohydraulics *see* Newtonian Fluids and Thermohydraulics

Toda Lattices *Y B Suris* 235

Toeplitz Determinants and Statistical Mechanics *E L Basor* 244

Tomita–Takesaki Modular Theory *S J Summers* 251

Topological Defects and Their Homotopy Classification *T W B Kibble* 257

Topological Gravity, Two-Dimensional *T Eguchi* 264

Topological Knot Theory and Macroscopic Physics *L Boi* 271

Topological Quantum Field Theory: Overview *J M F Labastida and C Lozano* 278

Topological Sigma Models *D Birmingham* 290

Turbulence Theories *R M S Rosa* 295

Twistor Theory: Some Applications *L Mason* 303

Twistors *K P Tod* 311

Two-Dimensional Conformal Field Theory and Vertex Operator Algebras *M R Gaberdiel* 317

Two-Dimensional Ising Model *B M McCoy* 322

Two-Dimensional Models *B Schroer* 328

U

Universality and Renormalization *M Lyubich* 343

V

Variational Methods in Turbulence *F H Busse* 351

Variational Techniques for Ginzburg–Landau Energies *S Serfaty* 355

Variational Techniques for Microstructures *G Dolzmann* 363

Vertex Operator Algebras *see* Two-Dimensional Conformal Field Theory and Vertex Operator Algebras

Viscous Incompressible Fluids: Mathematical Theory *J G Heywood* 369

von Neumann Algebras: Introduction, Modular Theory, and Classification Theory *V S Sunder* 379

von Neumann Algebras: Subfactor Theory *Y Kawahigashi* 385

Vortex Dynamics *M Nitsche* 390

Vortices *see* Abelian Higgs Vortices: Point-Vortex Dynamics

W

Wave Equations and Diffraction *M E Taylor* 401

Wavelets: Application to Turbulence *M Farge and K Schneider* 408

Wavelets: Applications *M Yamada* 420

Wavelets: Mathematical Theory *K Schneider and M Farge* 426

WDVV Equations and Frobenius Manifolds *B Dubrovin* 438

Weakly Coupled Oscillators *E M Izhikevich and Y Kuramoto* 448

Wheeler–De Witt Theory *J Maharana* 453

Wightman Axioms *see* Axiomatic Quantum Field Theory

Wulff Droplets *S Shlosman* 462

Y

Yang–Baxter Equations *J H H Perk and H Au-Yang* 465

INDEX 475

Stability of Flows

S Friedlander, University of Illinois-Chicago,
Chicago, IL, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

This article gives a brief discussion of a topic with an enormous literature, namely the stability/instability of fluid flows. Following the seminal observations and experiments of Reynolds in 1883, the issue of stability of a fluid flow became one of the central problems in fluid dynamics: stable flows are robust under inevitable disturbances in the environment, while unstable flows may break up, sometimes rapidly. These possibilities were demonstrated in a relatively simple experiment where flow in a pipe is examined at increasing speeds. As a dimensionless parameter (now known as the Reynolds number) increases, the flow completely changes its nature from a stable flow to a completely different regime that is irregular in space and time. Reynolds called this “turbulence” and observed that the transition from the simple flow to the chaotic flow was caused by the phenomenon of instability.

Even though the topic has been the subject of intense study over more than a century, Reynolds experiment is still not fully explained by current theory. Although there is no rigorous proof of stability of the simple flow (known as Poiseuille flow in a circular pipe), analytical and numerical investigations of the equations suggest theoretical stability for all Reynolds numbers. However, experiments show instability for sufficiently large Reynolds numbers. A plausible explanation for this phenomenon is the instability of such flows with respect to small but finite disturbances combined with their stability to infinitesimal disturbances.

The issue of fluid stability, in contexts much more complex than the fundamental experiment of Reynolds, arises in a multitude of branches of science, including engineering, physics, astrophysics, oceanography, and meteorology. It is far beyond the scope of this short article to even touch upon most of the extensive literature. In the bibliography we list just a few of the substantive books where classical results can be found (Chandrasekhar 1961, Drazin and Reid 1981, Gershuni and Zhukovitiskii 1976, Joseph 1976, Lin 1967, Swinney and Gollub 1985). Recent extensive bibliographies on mathematical aspects of fluid instability are given in several articles in the *Handbook of Mathematical Fluid Dynamics*

(Friedlander and Serre 2003) and the compendium of articles on hydrodynamics and nonlinear instabilities in Godreche and Maneville (1998).

The Equations of Motion

The Navier–Stokes equations for the motion of an incompressible, constant density, viscous fluid are

$$\frac{\partial \mathbf{q}}{\partial t} + (\mathbf{q} \cdot \nabla) \mathbf{q} = -\frac{1}{\rho} \nabla P + \nu \nabla^2 \mathbf{q} \quad [1a]$$

$$\operatorname{div} \mathbf{q} = 0 \quad [1b]$$

where $\mathbf{q}(\mathbf{x}, t)$ denotes the velocity vector, $P(\mathbf{x}, t)$ the pressure, and the constants ρ and ν are the density and kinematic viscosity, respectively. This system is considered in three (or sometimes two) spatial dimensions with a specified initial velocity field

$$\mathbf{q}(\mathbf{x}, 0) = \mathbf{q}_0(\mathbf{x}) \quad [1c]$$

and physically appropriate boundary conditions: for example, zero velocity on a rigid boundary, or periodicity conditions for flow on a torus. This nonlinear system of partial differential equations (PDEs) has proved to be remarkably challenging, and in three dimensions the fundamental issues of existence and uniqueness of physically reasonable solutions are still open problems.

It is often useful to consider the Navier–Stokes equations in nondimensional form by scaling the velocity and length by some intrinsic scale in the problem, for example, in Reynolds’ experiment by the mean speed U and the diameter of the pipe d . This leads to the nondimensional equations

$$\frac{\partial \mathbf{q}}{\partial t} + (\mathbf{q} \cdot \nabla) \mathbf{q} = -\nabla P + \frac{1}{R} \nabla^2 \mathbf{q} \quad [2a]$$

$$\operatorname{div} \mathbf{q} = 0 \quad [2b]$$

where the Reynolds number R is

$$R = Ud/\nu \quad [3]$$

In many situations, the size of R has a crucial influence on stability. Roughly speaking, when R is small the flow is very sluggish and likely to be stable. However, the effects of viscosity are actually very complicated and not only is viscosity able to smooth and stabilize fluid motions, sometimes it actually also destroys and destabilizes flows.

The Euler equations, which predate the Navier–Stokes equations by many decades, neglect the effects of viscosity and are obtained from [1a] by setting the viscosity parameter ν to zero. Since this

removes the highest-derivative term from the equations, the nature of the Euler equations is fundamentally different from that of the Navier–Stokes equations and the limit of vanishing viscosity (or infinite Reynolds number) is a very singular limit. Since all real fluids are at least very weakly viscous, it could be argued that only the the Navier–Stokes equations are physically relevant. However, many important physical phenomena, such as turbulence, involve flows at very high Reynolds numbers (10^4 or higher). Hence, an understanding of turbulence is likely to involve the asymptotics of the Navier–Stokes equations as $R \rightarrow \infty$. The first step towards the construction of such asymptotics is the study of inviscid fluids governed by the Euler equations:

$$\frac{\partial \mathbf{q}}{\partial t} + (\mathbf{q} \cdot \nabla) \mathbf{q} = -\nabla P \quad [4a]$$

$$\operatorname{div} \mathbf{q} = 0 \quad [4b]$$

Stability issues for the Euler equations are in many respects distinct from those of the Navier–Stokes equations and in this article we will briefly touch upon stability results for both systems.

Comments on Some “Classical” Instabilities

To illustrate the complexity of the structure of instabilities that can arise in the Navier–Stokes equations, we mention one classical example, namely the centrifugal instabilities called Taylor–Couette instabilities. Consider a fluid between two concentric cylinders rotating with different angular velocities. If the inner cylinder rotates sufficiently faster than the outer one, the centrifugal force is stronger on inside particles than outside particles and a disturbance which exchanges the radial position of particles is enhanced, that is, the configuration is unstable. As the angular velocity of the inner cylinder is increased above a certain critical rate, the instability is manifested in a series of small toroidal (Taylor) vortices that fill the space between the cylinders. There follows a hierarchy of successive instabilities: azimuthal traveling waves, twisting regimes, and quasiperiodic regimes until chaotic solutions appear. Such a sequence of bifurcations is a scenario for a transition to turbulence postulated by Ruelle–Takens. Details concerning bifurcation theory and fluid behavior can be found in the book of Chossat and Iooss (1994).

We note that phenomena of successive bifurcations connected with loss of stability, such as regimes of Taylor–Couette instabilities, occur at

moderately large Reynolds numbers. Fully developed turbulence is a phenomenon associated with very high Reynolds numbers. These are parameter regimes basically inaccessible in current numerical investigations of the Navier–Stokes equations and turbulent models. The Euler equations lie at the limit as $R \rightarrow \infty$. It is an interesting observation that results at the limit of infinite Reynolds number are sometimes also applicable and consistent with experiments for flows with only moderate Reynolds number.

There is a huge diversity of forces that couple with fluid motion to produce instability. We will merely mention a few of these which an interested reader could pursue in consultation with texts listed in the “Further reading” section and references therein.

1. The so-called Bénard problem of convective instability concerns a horizontal layer of fluid between parallel plates and subject to a temperature gradient. The governing equations are the Navier–Stokes equation for a nonconstant density fluid and the heat equation. In this problem, the critical parameter governing the onset of instability is called the Rayleigh number. The patterns that can develop as a result of instability are strongly influenced by the boundary conditions in the horizontal coordinates. With lattice type conditions, bifurcating solutions include rolls, rectangles, and hexagons. Convection rolls are themselves subject to secondary instabilities that may break the translation symmetry and deform the rolls into meandering shapes. Further refinements of convective instabilities include doubly diffusive convection, where the density depends on concentration as well as temperature. Competition between stabilizing diffusivity and destabilizing diffusivity can lead to the so-called “salt-finger” instabilities.
2. Of considerable interest in astrophysics and plasma physics are the instabilities that occur in electrically conducting fluids. Here the fluid equations are coupled with Maxwell’s equations. Much work has been done on the topic of magnetohydrodynamical (MHD) stability, which was developed to address various important physical issues such as thermonuclear fusion, stellar and planetary interiors, and dynamo theory. For example, dynamo theory addresses the issue of how a magnetic field can be generated and sustained by the motion of an electrically conducting fluid. In the simplest scenario, the fluid motion is assumed to be a given divergence-free vector field and the study of

the instabilities that may occur in the evolution of the magnetic field is called the kinematic dynamo problem. This gives rise to interesting problems in dynamical systems and actually is closely analogous to the topic of vorticity generation in the three-dimensional (3D) fluid equations in the absence of MHD effects.

In the next section we discuss certain mathematical results that have been rigorously proved for particular problems in the stability of fluid flows. We restrict our attention to the “basic” equations, that is, [2a] and [2b], [4a] and [4b], observing that even in rather simple configurations there are still more open problems than precise rigorous results.

The Navier–Stokes Equations: Mathematical Definitions of Stability/Instability

Instability occurs when there is some disturbance of the internal or external forces acting on the fluid and, loosely speaking, the question of stability or instability considers whether there exist disturbances that grow with time. There are many mathematical definitions of stability of a solution to a PDE. Most of these definitions are closely related but they may not be equivalent. Because of the distinctly different nature of the Navier–Stokes equations for a viscous fluid and the Euler equations for an inviscid fluid, we will adopt somewhat different precise definitions of stability for the two systems of PDEs. Both definitions are related to the concept known as Lyapunov stability. A steady state described by a velocity field $U_0(\mathbf{x})$ is called Lyapunov stable if every state $q(\mathbf{x}, t)$ “close” to $U_0(\mathbf{x})$ at $t=0$ stays close for all $t > 0$. In mathematical terms, “closeness” is defined by considering metrics in a normed space X . While in finite-dimensional systems the choice of norm is not significant because all Banach norms are equivalent, in infinite-dimensional systems, such as a fluid configuration, this choice is crucial. The point was emphasized by Yudovich (1989) and it is a version of the definition of stability given in this book that we will adopt in connection with the parabolic Navier–Stokes equations.

Definitions for a General Nonlinear Evolution Equation

Consider an evolution equation for $u(\mathbf{x}, t)$ whose phase space is a Banach space X :

$$\frac{\partial u}{\partial t} = Lu + N(u, u)$$

We assume that if the initial value $u(\mathbf{x}, 0) \in X$ is given, the future evolution $u(\mathbf{x}, t), t > 0$, of the equation is uniquely defined (at least for sufficiently small initial data). Without loss of generality, we can assume that zero is a steady state.

We define a version of Lyapunov (nonlinear) stability and its converse instability.

Definition 1 Let (X, Z) be a pair of Banach spaces. The zero steady state is called (X, Z) nonlinearly stable if, no matter how small $\epsilon > 0$, there exists $\delta > 0$ so that $u(\mathbf{x}, 0) \in X$ and

$$\|u(\mathbf{x}, 0)\|_Z < \delta$$

imply the following two assertions:

- (i) there exists a global in time solution such that $u(\mathbf{x}, t) \in ([0, \infty); X)$;
- (ii) $\|u(\mathbf{x}, t)\|_Z < \epsilon$ for a.e. $t \in [0, \infty)$.

The zero state is called nonlinearly unstable if either of the above assertions is violated. We note that under this strong definition of stability, loss of existence of a solution is a particular case of instability. The concept of existence that we will invoke in considering the Navier–Stokes equations is the existence of “mild” solutions introduced by Kato and Fujita (1962). Local-in-time existence of mild solutions is known in $X = L^q$ for $q \geq n$, where n denotes the space dimension. (L^q denotes the usual Lebesgue space).

We now state two theorems for the Navier–Stokes equations [2a] and [2b]. The theorems are valid in any space dimension n and in finite or infinite domains. Of course, the most physically relevant cases are $n=3$ or 2 . Both theorems relate properties of the spectrum of the linearized Navier–Stokes equations to stability or instability of the full nonlinear system. Let $U_0(\mathbf{x}), P_0(\mathbf{x})$ be a steady state flow:

$$(U_0 \cdot \nabla)U_0 = -\nabla P_0 + \frac{1}{R}\nabla^2 U_0 + \frac{1}{R}F \quad [5a]$$

$$\nabla \cdot U_0 = 0 \quad [5b]$$

where $U_0 \in C^\infty$ vanishes on the boundary of the domain D and F is a suitable external force. We write [2a] and [2b] in perturbation form as

$$q(\mathbf{x}, t) = U_0(\mathbf{x}) + u(\mathbf{x}, t) \quad [6]$$

where

$$\frac{\partial u}{\partial t} = L_{NS}u + N(u, u) \quad [7a]$$

$$\nabla \cdot u = 0 \quad [7b]$$

with

$$L_{NS}u \equiv -(U_0 \cdot \nabla)u - (u \cdot \nabla)U_0 + \frac{1}{R}\nabla^2 u - \nabla P_1 \quad [8]$$

$$N(u, u) \equiv -(u \cdot \nabla)u - \nabla P_2 \quad [9]$$

Here P_1 and P_2 are, respectively, the portions of the pressure required to ensure that $L_{NS}u$ and $N(u, u)$ remain divergence free. The operators L_{NS} and N act on the space of divergence-free vector-valued functions in the closure of the Sobolev space $W^{s,p}$ that vanish on the boundary of D .

We note that the spectrum of the elliptic linear operator L_{NS} with appropriate boundary conditions in a bounded domain is purely discrete: that is, it consists of a countable number of eigenvalues of finite multiplicity with the sole limit point being at infinity.

Theorem 2 (Nonlinear instability). *Let $1 < p < \infty$ be arbitrary. Suppose that the operator L_{NS} over L^p has spectrum in the right half of the complex plane. Then the flow $U_0(x)$ is (L^q, L^p) nonlinearly unstable for any $q > \max(p, n)$.*

Theorem 3 (Asymptotic Lyapunov stability). *Let $q > n$ be arbitrary. Assume that the operator L_{NS} over L^q has spectrum confined to the left half of the complex plane. Then the flow $U_0(x)$ is (L^q, L^q) nonlinearly stable.*

A recent proof of these theorems is given in Friedlander *et al.* (2006) using a bootstrap type argument. In Theorem 2, the space $L^q, q > n$, is used as an auxiliary space in which the norm of the nonlinear term is controlled, while the final instability result is proved in L^p for $p \in (1, \infty)$. We note that this includes the most physically relevant case of instability in the L^2 energy norm. An earlier proof of the theorems under the restriction $p \geq n$ was given by Yudovich (1989).

To apply Theorem 2 or 3 to conclude nonlinear instability or stability of a given flow U_0 , it is necessary to have information concerning the spectrum of the linear operator L_{NS} . Obtaining such information has been the goal of much of the literature concerning fluid stability (see the bibliography and the references therein). However, except in the case of some relatively simple flows, the eigenvalues of L_{NS} have not yet been calculated explicitly. Perhaps the example that is the most tractable is plane parallel shear flows. Here the eigenvalue problem is governed by an ordinary differential equation (ODE) known as the Orr–Sommerfeld equation, which has been the subject of

extensive analytical and numerical investigations. Consider the parallel flow $U_0 = (U(z), 0, 0)$ in the strip $-1 \leq z \leq 1$. For disturbances of the form

$$\phi(z) e^{i(k_1 x + k_2 y)} e^{\lambda t} \quad [10]$$

the eigenvalue λ is determined by the following equation with $k^2 = k_1^2 + k_2^2$:

$$\begin{aligned} \left(U - i \frac{\lambda}{k} \right) \left[\frac{d^2}{dz^2} - k^2 \right] \phi - U'' \phi \\ = \frac{1}{ikR} \left[\frac{d^2}{dz^2} - k^2 \right] \phi \end{aligned} \quad [11]$$

with boundary conditions $\phi = 0$ at $z = \pm 1$. We note that the discreteness of the spectrum is preserved if periodicity conditions are imposed in the (x, y) plane.

The complexity of the spectral problem [11] is apparent even for the simple case $U(z) = 1 - z^2$ (known as plane Poiseuille flow). Unstable eigenvalues exist but only in certain regions of (k, R) parameter space. There is a critical Reynolds number, $R_c = 5772$, below which $\text{Re } \lambda < 0$ for all wave numbers k . For $R > R_c$, instability occurs in a band of wave numbers and the thickness of this band shrinks to zero as $R \rightarrow \infty$ (i.e., the inviscid limit). Hence, Poiseuille flow with $R < R_c$ can be considered as an example where the stability Theorem 3 can be applied, that is, the flow is nonlinearly stable to infinitesimal disturbances. However, extremely careful experiments are needed to obtain agreement with the theoretical value of $R_c = 5772$. Rather it is more usual in an experiment with $R \sim 2000$ that the flow exhibits instability in the form of streamwise streaks that appear near the walls. These structures do not look like traveling waves of the form given by expression [10], rather they are finite-amplitude effects of nonmodal growth. Such linear growth of disturbances, along with energy growth and pseudospectra have recently been investigated extensively.

An example where Theorem 3, proving nonlinear instability, can be applied is the so-called Kolmogorov flow. This is also a shear flow with the spectral problem for the linearized operator given by eqn [11]. In this example, the profile is oscillatory in z with $U(z) = \sin mz$. In an elegant paper, Meshalkin and Sinai (1961) used continued fractions to prove the existence of a real unstable positive eigenvalue. It is interesting, and in some sense surprising, that the particular case of sinusoidal profiles leads to a nonconstant-coefficient eigenvalue problem, where it is possible to construct in explicit form the transcendental characteristic equation that relates the eigenvalues λ and the wave numbers. Usually,

this can be done only for constant-coefficient equations. In the case $U(z) = \sin mz$, a Fourier series representation for the eigenfunctions leads to a tridiagonal infinite matrix for the algebraic system satisfied by the Fourier coefficients. This is amenable to examination using continued fractions. Analysis of the characteristic equation shows that there exist real eigenvalues $\lambda > 0$ provided R is larger than some critical value for each wave number k with $k^2 < m^2$.

The Euler Equation: Linear and Nonlinear Stability/Instability

We conclude this brief article with some discussion of instabilities in the inviscid Euler equations whose existence is likely to be important as a “trigger” for the development of instabilities in high-Reynolds-number viscous flows. As we mentioned, the Euler equations are very different from the Navier–Stokes equations in their mathematical structure. The Euler equations are degenerate and nonelliptic. As such, the spectrum of the linearized operator L_E is not amenable to standard spectral theory of elliptic operators. For example, unlike the Navier–Stokes operator, the spectrum of L_E is not purely discrete even in bounded domains. To define L_E we consider a steady Euler flow $\{U_0(x), P_0(x)\}$, where

$$U_0 \cdot \nabla U_0 = -\nabla P_0 \quad [12a]$$

$$\nabla \cdot U_0 = 0 \quad [12b]$$

We assume that $U_0 \in C^\infty$. For the Euler equations, appropriate boundary conditions include zero normal component of U_0 on a rigid boundary, or periodicity conditions (i.e., flow on a torus) or suitable decay at infinity in an unbounded domain. The theorems that we will be describing have been proved mainly in the cases of the second and third conditions stated above. There are many classes of vector fields $U_0(x)$, in two and three dimensions, that satisfy [12a] and [12b]. We write [4a] and [4b] in perturbation form as

$$q(x, t) = U_0(x) + u(x, t) \quad [13]$$

with

$$\frac{\partial u}{\partial t} = L_E u + N(u, u) \quad [14a]$$

$$\nabla \cdot u = 0 \quad [14b]$$

Here

$$L_E u \equiv -(U_0 \cdot \nabla) u - (u \cdot \nabla) U_0 - \nabla P_1 \quad [15]$$

$$N(u, u) \equiv -(u \cdot \nabla) u - \nabla P_2 \quad [16]$$

Linear (spectral) instability of a steady Euler flow $U_0(x)$ concerns the structure of the spectrum of L_E . Assuming $U_0 \in C^\infty(T^n)$, the linear equation

$$\frac{\partial u}{\partial t} = L_E u, \quad \nabla \cdot u = 0 \quad [17]$$

defines a strongly continuous group in every Sobolev space $W^{s,p}$ with generator L_E . We denote this group by $\exp\{L_E t\}$. For the issue of spectral instability of the Euler equation it proves useful to study not only the spectrum of L_E but also the spectrum of the evolution operator $\exp\{L_E t\}$. This permits the development of an explicit formula for the growth rate of a small perturbation due to the essential (or continuous) spectrum. It was proved by Vishik (1996) that a quantity Λ , referred to as a “fluid Lyapunov exponent” gives the maximum growth rate of the essential spectrum of $\exp\{L_E t\}$. This quantity is obtained by computing the exponential growth rate of a certain vector that satisfies a specific system of ODEs over the trajectories of the flow $U_0(x)$. This proves to be an effective mechanism for detecting instabilities in the essential spectrum which result due to high-spatial-frequency perturbations. For example, for this reason any flow $U_0(x)$ with a hyperbolic fixed point is linearly unstable with growth in the sense of the L^2 -norm. In two dimensions, Λ is equal to the maximal classical Lyapunov exponent (i.e., the exponential growth of a tangent vector over the ODE $\dot{x} = U_0(x)$). In three dimensions, the existence of a nonzero classical Lyapunov exponent implies that $\Lambda > 0$. However, in three dimensions there are also examples where the classical Lyapunov exponent is zero and yet $\Lambda > 0$. We note that the delicate issue of the unstable essential spectrum is strongly dependent on the function space for the perturbations and that Λ , for a given U_0 , will vary with this function space. More details and examples of instabilities in the essential spectrum can be found in references in the bibliography.

In contrast with instabilities in the essential spectrum, the existence of discrete unstable eigenvalues is independent of the norm in which growth is measured. From this point of view, such instabilities can be considered as “strong.” However, for most flows $U_0(x)$ we do not know the existence of such unstable eigenvalues. For fully 3D flows there are no examples, to our knowledge, where such unstable eigenvalues have been proved to exist for flows with standard metrics. The case that has received the most attention in the literature is the “relatively simple” case of plane parallel shear flow. The eigenvalue problem is governed by the Rayleigh

equation (which is the inviscid version of the Orr–Sommerfeld equation [11]):

$$\left(U - \frac{i\lambda}{k}\right) \left[\frac{d^2}{dz^2} - k^2\right] \phi - U''\phi = 0$$

$$\phi = 0 \quad \text{at} \quad z = \pm 1 \quad [18]$$

The celebrated Rayleigh stability criterion says that a sufficient condition for the eigenvalues λ to be pure imaginary is the absence of an inflection point in the shear profile $U(z)$. It is more difficult to prove the converse; however, there have been several recent results that show that oscillating profiles indeed produce unstable eigenvalues. For example, if $U(z) = \sin mz$ the continued fraction proof of Meshalkin and Sinai can be adapted to exhibit the full unstable spectrum for [18]. We note the “fluid Lyapunov exponent” Λ is zero for all shear flows; thus the only way the unstable spectrum can be nonempty for shear flows is via discrete unstable eigenvalues.

As we have discussed, it is possible to show that many classes of steady Euler flows are linearly unstable, either due to a nonempty unstable essential spectrum (i.e., cases where $\Lambda > 0$) or due to unstable eigenvalues or possibly for both reasons. It is natural to ask what this means about the stability/instability of the full nonlinear Euler equations [14]–[16]. The issue of nonlinear stability is complex and there are several natural precise definitions of nonlinear stability and its converse instability.

One definition is to consider nonlinear stability in the energy norm L^2 and the enstrophy norm H^1 , which are natural function spaces to measure growth of disturbances but are not “correct” spaces for the Euler equations in terms of proven properties of existence and uniqueness of solutions to the nonlinear equation. Falling under this definition is the most frequently employed method to prove nonlinear stability, which is an elegant technique developed by Arnol’d (cf. Arnol’d and Khesin (1998) and references therein). This is based on the existence of the so-called energy-Casimirs. The vorticity $\text{curl } q$ is transported by the motion of the fluid so that at time t it is obtained from the vorticity at time $t=0$ by a volume-preserving diffeomorphism. In the terminology of Arnol’d, the velocity fields obtained in this manner at any two times are called isovortical. For a given field $U_0(x)$, the class of isovortical fields is an infinite-dimensional manifold M , which is the orbit of the group of volume-preserving diffeomorphisms in the space of divergence-free vector fields. The steady flows are exactly the critical points of the energy functional E restricted to M . If a critical point is a

strict local maximum or minimum of E , then the steady flow is nonlinearly stable in the space J_1 of divergence-free vectors $u(x, t)$ (satisfying the boundary conditions) that have finite norm,

$$\|u\|_{J_1} \equiv \|u\|_{L^2} + \|\text{curl } u\|_{L^2} \quad [19]$$

This theory can be applied, for example, to show that any shear flow with no inflection points in the profile $U(z)$ is nonlinearly unstable in the function space J_1 , that is, the classical Rayleigh criterion implies not only spectral stability but also nonlinear stability.

We note that Arnol’d’s stability method cannot be applied to the Euler equations in three dimensions because the second variation of the energy defined on the tangent space to M is never definite at a critical point $U_0(x)$. This result is suggestive, but does not prove, that most Euler flows in three dimensions are nonlinearly unstable in the Arnol’d sense. To quote Arnol’d, in the context of the Euler equations “there appear to be an infinitely great number of unstable configurations.”

In recent years, there have been a number of results concerning nonlinear instability for the Euler equation. Most of these results prove nonlinear instability under certain assumptions on the structure of the spectrum of the linearized Euler operator. To date, none of the approaches prove the definitive result that in general linear instability implies nonlinear instability. As we have remarked, this is a much more delicate issue for Euler than for Navier–Stokes because of the existence, for a generic Euler flow, of a nonempty essential unstable spectrum. To give a flavor of the mathematical treatment of nonlinear instability for the Euler equations, we present one recent result and refer the interested reader to articles listed in the “Further reading” section for further results and discussions.

In the context of Euler equations in two dimensions, we adopt the following definition of Lyapunov stability.

Definition 4 An equilibrium solution $U_0(x)$ is called Lyapunov stable if for every $\varepsilon > 0$ there exists $\delta > 0$ so that for any divergence-free vector $u(x, 0) \in W^{1+s,p}$, $s > 2/p$, such that $\|u(x, 0)\|_{L^2} < \delta$ the unique solution $u(x, t)$ to [14]–[16] satisfies

$$\|u(x, t)\|_{L^2} < \varepsilon \quad \text{for } t \in [0, \infty)$$

We note that we require the initial value $u(x, 0)$ to be in the Sobolev space $W^{1+s,p}$, $s > p/2$, since it is known that the 2D Euler equations are globally in time well posed in this function space.

Definition 5 Any steady flow $U_0(x)$ for which the conditions of Definition 4 are violated is called nonlinearly unstable in L^2 .

Observe that the open issues (in three dimensions) of nonuniqueness or nonexistence of solutions to [14]–[16] would, under Definition 5, be scenarios for instability.

Theorem 6 (Nonlinear instability for 2D Euler flows). *Let $U_0(x) \in C^\infty(T^2)$ be satisfy [12]. Let Λ be the maximal Lyapunov exponent to the ODE $\dot{x} = U_0(x)$. Assume that there exists an eigenvalue λ in the L^2 spectrum of the linear operator L_E given by [15] with $\operatorname{Re} \lambda > \Lambda$. Then in the sense of Definition 5, $U_0(x)$ is Lyapunov unstable with respect to growth in the L^2 -norm.*

The proof of this result is given in Vishik and Friedlander (2003) and uses a so-called “bootstrap” argument whose origins can be found in references in that article. We remark that the above result gives nonlinear instability with respect to growth of the energy of a perturbation which seems to be a physically reasonable measure of instability.

In order to apply Theorem 6 to a specific 2D flow it is necessary to know that the linear operator L_E has an eigenvalue with $\operatorname{Re} \lambda > \Lambda$. As we have discussed, such knowledge is lacking for a generic flow $U_0(x)$. Once again, we turn to shear flows. As we noted $\Lambda = 0$ for shear flows, any shear profile for which unstable eigenvalues have been proved to exist provides an example of nonlinear instability with respect to growth in the energy.

We conclude with the observation that it is tempting to speculate that, given the complexity of flows in three dimensions, most, if not all, such inviscid flows are nonlinearly unstable. It is clear from the concept of the fluid Lyapunov exponent that stretching in a flow is associated with instabilities and there are more mechanisms for stretching in three, as opposed to two, dimensions. However, to date there are virtually no mathematical results for the nonlinear stability problem for fully 3D flows and many challenging issues remain entirely open.

Acknowledgments

The author is very grateful to IHES and ENS-Cachan for their hospitality during the writing of

this paper. She thanks Misha Vishik for much helpful advice.

The work is partially supported by NSF grant DMS-0202767.

See also: Compressible Flows: Mathematical Theory; Incompressible Euler Equations: Mathematical Theory; Magnetohydrodynamics; Newtonian Fluids and Thermohydraulics; Non-Newtonian Fluids; Topological Knot Theory and Macroscopic Physics.

Further Reading

- Arnol'd VI and Khesin B (1998) *Topological Methods in Hydrodynamics*. New York: Springer.
- Chandrasekhar S (1961) *Hydrodynamic and Hydromagnetic Stability*. Oxford: Oxford University Press.
- Chossat P and Iooss G (1994) *The Couette–Taylor Problem*. Berlin: Springer.
- Drazin PG and Reid WH (1981) *Hydrodynamic Stability*. Cambridge: Cambridge University Press.
- Friedlander S, Pavlovic N, and Shvydkoy R (2006) Nonlinear instability for the Navier–Stokes equations (to appear in *Communications in Mathematical Physics*).
- Friedlander S and Serre D (eds.) (2003) *Handbook of Mathematical Fluid Dynamics*, vol. 2. Amsterdam: Elsevier.
- Friedlander S and Yudovich VI (1999) Instabilities in fluid motion. *Notices of the American Mathematical Society* 46: 1358–1367.
- Gershuni GZ and Zhukovitiskii EM (1976) *Convective Instability of Incompressible Fluids*. Jerusalem: Keter Publishing House.
- Godreche C and Manneville P (eds.) (1998) *Hydrodynamics and Nonlinear Instabilities*. Cambridge: Cambridge University Press.
- Joseph DD (1976) *Stability of Fluid Motions*, 2 vols. Berlin: Springer.
- Kato T and Fujita H (1962) On the nonstationary Navier–Stokes system. *Rend. Sem. Mat. Univ. Padova* 32: 243–260.
- Lin CC (1967) *The Theory of Hydrodynamic Stability*. Cambridge: Cambridge University Press.
- Meshalkin LD and Sinai IaG (1961) Investigation of stability for a system of equations describing the plane movement of an incompressible viscous liquid. *App. Math. Mech.* 25: 1700–1705.
- Swinney H and Gollub L (eds.) (1985) *Hydrodynamic Instabilities and Transition to Turbulence*. New York: Springer.
- Vishik MM (1996) Spectrum of small oscillations of an ideal fluid and Lyapunov exponents. *Journal de Mathématiques Pures et Appliquées* 75: 531–557.
- Vishik M and Friedlander S (2003) Nonlinear instability in 2 dimensional ideal fluids: the case of a dominant eigenvalue. *Communications in Mathematical Physics* 243: 261–273.
- Yudovich VI (1989, US translation) *Linearization Method in Hydrodynamical Stability Theory*, Transl. Math. Monog. vol. 74. Providence: American Mathematical Society.

Stability of Matter

J P Solovej, University of Copenhagen, Copenhagen, Denmark

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The theorem on stability of matter is one of the most celebrated results in mathematical physics. It is one of the rare cases where a result of such great importance to our understanding of the world around us appeared first in a completely rigorous formulation.

Issues of stability are, of course, extremely important in physics. One of the major triumphs of the theory of quantum mechanics is the explanation it gives of the stability of the hydrogen atom (and the complete description of its spectrum). Quantum mechanics or, more precisely, the uncertainty principle explains not only the stability of tiny microscopic objects, but also the stability of gigantic stellar objects such as white dwarfs. Chandrasekhar's famous theory on the stability of white dwarfs required, however, not only the usual uncertainty principle, but also the Pauli exclusion principle for the fermionic electrons.

Whereas both the stability of atoms and the stability of white dwarfs were early triumphs of quantum mechanics, it, surprisingly, took nearly 40 years before the question of stability of everyday macroscopic objects was even raised (Fisher and Ruelle 1966). The rigorous answer to the question came shortly thereafter in what came to be known as the "theorem on stability of matter" proved first by Dyson and Lenard (1967).

Both the stability of hydrogen and the stability of white dwarfs simply mean that the total energy of the system cannot be arbitrarily negative. If there were no such lower bound to the energy, one would have a system from which it would be possible, in principle, to extract an infinite amount of energy. One often refers to this kind of stability as *stability of the first kind*.

Stability of matter is somewhat different. Stability of the first kind for atoms generalizes, as noted later, to objects of macroscopic size. The question arises as to how the lowest possible energy depends on the size or, more precisely, on the (macroscopic) number of particles in the object. Stability of matter in its precise mathematical formulation is the requirement that the lowest possible energy depends at most linearly on the number of particles. Put differently, the lowest possible energy calculated per particle

cannot be arbitrarily negative as the number of particles increases. This is often referred to as "stability of the second kind." If stability of the second kind does not hold, one would be able to extract an arbitrarily large amount of energy by adding a single atomic particle to a sufficiently large macroscopic object.

A perhaps more intuitive notion of stability is related to the volume occupied by a macroscopic object. More precisely, the volume of the object, when its total energy is close to the lowest possible energy, grows at least linearly in the number of particles. This volume dependence is a fairly simple consequence of stability of matter as formulated above.

The first mention of stability of the second kind for a charged system is perhaps by Onsager (1939), who studied a system of charged classical particles with a hard core and proved the stability of the second kind. The proof of stability of matter by Dyson and Lenard, which does not rely on any hard-core assumption, but rather on the properties of fermionic quantum particles, used results from Onsager's paper.

The real relevance of the notion of stability of the second kind was first realized by Fisher and Ruelle (1966) in an attempt to understand the thermodynamic properties of matter and to give meaning to thermodynamic quantities such as the energy density (energy per volume). Stability of matter is a necessary ingredient in explaining the existence of thermodynamics, that is, that the energy per volume has a well-defined limit as the volume and number of particles tend to infinity, with the ratio (i.e., the density of particles) kept fixed. The existence of this limit is, however, not just a simple consequence of stability of matter. The existence of the thermodynamic limit for ordinary charged matter was proved rigorously by Lieb and Lebowitz (1972) using the result on stability of matter as an input.

After the original proof of stability of matter by Dyson and Lenard, several other proofs were given (see, e.g., reviews by Lieb (1976, 1990, 2004) for detailed references). Lieb and Thirring (1975) in particular presented an elegant and simple proof relying on an uncertainty principle for fermions. As explained in a later section, the best mathematical formulation of the usual uncertainty principle is in terms of a Sobolev inequality. The method of Lieb and Thirring is related to a Sobolev type inequality for antisymmetric functions. The Lieb–Thirring inequality is discussed later. The proof by Dyson

and Lenard gave a very poor bound on the lowest possible energy per particle. The proof by Lieb and Thirring gave a much more realistic bound on this quantity (see below). Two proofs of stability of matter will be sketched here. Both proofs rely on the Lieb–Thirring inequality. The first proof described is mathematically simple to explain, whereas the second proof (Lieb–Thirring) is based on the Thomas–Fermi theory. It is mathematically somewhat more involved but, from a physical point of view, more intuitive.

As in the case of white dwarfs, stability of matter relies on the fermionic property of electrons. Dyson (1967) proved that the stability of the second kind fails if we ignore the Pauli exclusion principle. In physics textbooks, the importance of the Pauli exclusion principle for the stability of white dwarfs is often emphasized. Its importance for the stability of everything around us is usually ignored.

As mentioned above the result on stability of matter appeared from the beginning as a completely rigorously proved theorem. In contrast, the stability of white dwarfs was only derived rigorously by Lieb and Thirring (1984) and Lieb and Yau (1987) over 50 years after the original work of Chandrasekhar.

The original formulation of stability of matter, which is given in the next section, dealt with charged matter consisting of electrons and nuclei interacting only through electrostatic interactions and being described by nonrelativistic quantum mechanics. Over the years, many generalizations of stability of matter have been derived in order to include relativistic effects and electromagnetic interactions. Some of these generalizations will be discussed in this article. A complete understanding of stability of matter in quantum electrodynamics (QED) does not exist as yet, which is intimately related to the fact that this theory still awaits a mathematically satisfactory formulation.

The Formulation of Stability of Matter

Consider K nuclei with nuclear charges $z_1, \dots, z_K > 0$ at positions $r_1, \dots, r_K \in \mathbb{R}^3$, and N electrons with charges -1 (this amounts to a choice of units) at positions $x_1, \dots, x_N \in \mathbb{R}^3$. In order to discuss stability, it turns out that one can consider the nuclei as fixed in space, whereas the electrons are dynamic. More precisely, this means that the kinetic energy of the nuclei is ignored. It is important to realize that if stability holds for static nuclei, it also holds for dynamic nuclei. This is simply because the kinetic energy is positive, so that the effect of ignoring it is to lower the total energy.

Since we consider only electrostatic interactions, the quantum Hamiltonian describing this system is

$$H_N = \sum_{i=1}^N T_i - \sum_{k=1}^K \sum_{i=1}^N \frac{z_k}{|x_i - r_k|} + \sum_{1 \leq i < j \leq N} \frac{1}{|x_i - x_j|} + \sum_{1 \leq k < \ell \leq K} \frac{z_k z_\ell}{|r_k - r_\ell|} \quad [1]$$

The kinetic energy operator T_i is (half) the Laplacian in the variable x_i , i.e., $T_i = -(1/2)\Delta_i$. Atomic units are used, where not only the electron charge is -1 , but the mass of the electron is also 1 and $\hbar = 1$. The unit of energy is then 2 Ry .

The Hamiltonian H_N depends on the parameters $z = (z_1, \dots, z_K)$ and $r = (r_1, \dots, r_K)$. It acts on the Hilbert space of fermionic, that is, antisymmetric wave functions. More precisely, the fermionic Hilbert space is

$$\mathcal{H}_N^F = \bigwedge^N L^2(\mathbb{R}^3; \mathbb{C}^2)$$

Here the target space is \mathbb{C}^2 , in order to describe spin-1/2 particles. One can, of course, also consider the Hamiltonian H_N on the full Hilbert space,

$$\mathcal{H}_N = \bigotimes_{i=1}^N L^2(\mathbb{R}^3; \mathbb{C}^2) = L^2(\mathbb{R}^{3N}; \mathbb{C}^{2^N})$$

of which \mathcal{H}_N^F is a subspace.

The quantity of interest is the ground-state energy

$$\begin{aligned} E^F(z, N, K) &= \inf_r \inf \text{spec}_{\mathcal{H}_N^F} H_N \\ &= \inf_r \inf \left\{ \langle \Psi, H_N \Psi \rangle \mid \Psi \right. \\ &\quad \left. \in \mathcal{H}_N^F \cap C^\infty(\mathbb{R}^{3N}; \mathbb{C}^{2^N}), \|\Psi\| = 1 \right\} \quad [2] \end{aligned}$$

and likewise for the ground-state energy $E(z, N, K)$ on the full space \mathcal{H}_N . Clearly, $E^F(z, N, K) \geq E(z, N, K)$. It turns out that the energy $E(z, N, K)$ is the same as one would get by restricting to symmetric functions instead of antisymmetric ones. Therefore, the energy $E(z, N, K)$ is often referred to as the lowest possible energy for bosonic particles.

The Hamiltonian H_N is an unbounded operator and we must discuss its domain to be able to talk about its spectrum. Also, it should be self-adjoint. It turns out that these questions are intimately related to stability. The operator H_N is well defined on smooth (i.e., C^∞) functions. Thus, the last definition of $E^F(z, N, K)$ in [2] is meaningful. If this ground-state energy is finite (i.e., not $-\infty$), then the Hamiltonian has an extension, the Friedrichs' extension, to a

self-adjoint operator with the property that the second equality in [2] holds.

In the definition of E^F , we have minimized over all the positions \mathbf{r} of the nuclei. Even though the nuclear dynamics is not considered, one is still interested in finding the lowest possible energy independent of where they are located.

Theorem 1 (Stability of the first kind). *For all N , K , and z , we have*

$$E(z, N, K) > -\infty$$

Theorem 2 (Stability of matter). *There exists a constant $C_{|z|} > 0$ depending only on $|z| = \max\{z_1, \dots, z_k\}$ such that*

$$E^F(z, N, K) \geq -C_{|z|}(N + K)$$

The constant $C_{|z|}$ bounds the binding energy per particle. In the case of hydrogen atoms, when $|z| = 1$, Dyson and Lenard arrived at a bound with $C_1 \approx 10^{14}$ Ry. Lieb and Thirring arrive at $C_1 \approx 5 = 10$ Ry. Since the binding energy of a single hydrogen atom is 1 Ry, it is easy to see that one must have $C_1 \geq 1/4$. Over the years, there have been some improvements on the estimated value of this constant in the theory of stability of matter.

That the Pauli exclusion principle, that is, the fermionic character of the electrons, is necessary for stability of matter is a consequence of the next theorem.

Theorem 3 (the $N^{5/3}$ law for bosons). *If $N = K$ and $z_1 = \dots = z_K = z > 0$, then there exist constants $C_{\pm} > 0$ depending on z such that*

$$-C_- N^{5/3} < E(z, N, N) < -C_+ N^{5/3}$$

It is the superlinear (exponent $5/3$) behavior in N of the upper bound that violates stability of matter. This upper bound was proved by Lieb (1979) by a fairly simple variational argument. The lower bound above, which shows that the exponent $5/3$ is optimal, was proved by Dyson and Lenard (1968) in their original paper on stability of matter.

This theorem leaves open the possibility that the stability of matter could be recovered by introducing finite nuclear masses. That this, indeed, is not the case was proved by Dyson (1967) by a complicated variational argument based on the Bogolubov pair theory for superfluid helium. We now add the kinetic energy $\sum_{k=1}^K -(1/2)\Delta_{r_k}$ of the nuclei (assuming, for simplicity, that they have the same mass as the electrons) to the Hamiltonian H_N and consider the case where $z_1 = z_2 = \dots = z_K = 1$. We denote the

ground-state energy over the space $L^2(\mathbb{R}^{3(N+K)})$ (ignoring spin) by $E(N, K)$. Then, Dyson proved that

$$\min_{N+K=M} \tilde{E}(N, K) \leq -CM^{7/5}$$

for some constant $C > 0$. It was later shown by Conlon *et al.* (1988) that the exponent $7/5$ is indeed optimal. Dyson (1967) made a conjecture for the precise asymptotic behavior of this energy. This conjecture, which was proved by Lieb and Solovej (2005) and Solovej (2004), is given in the next theorem.

Theorem 4 (Dyson's $7/5$ -law for the charged Bose gas).

$$\lim_{M \rightarrow \infty} \min_{N+K=M} \frac{\tilde{E}(N, K)}{M^{7/5}} = \inf \left\{ \frac{1}{2} \int |\nabla \phi|^2 - J \int \phi^{5/2} \mid \phi \geq 0, \int \phi^2 = 1 \right\} \quad [3]$$

where

$$J = \left(\frac{4}{\pi} \right)^{3/4} \frac{\Gamma(1/2)\Gamma(3/4)}{5\Gamma(5/4)}$$

Generalizations of Stability of Matter

Over the years, generalizations of stability of matter including relativistic effects and interactions with the electromagnetic field have been attempted. Since the relativistic Dirac operator is not bounded below, we cannot simply replace the standard nonrelativistic kinetic energy operator $T_j = -(1/2)\Delta_j$ by the free Dirac operator.

Relativistic effects have been included by considering the (pseudo) relativistic kinetic energy

$$T_j^{\text{Rel}} = \sqrt{-c^2 \Delta_j + c^4} - c^2$$

In the units used in this article, the physical value of the speed of light c is approximately 137 or, more precisely, the reciprocal of the fine-structure constant α .

For this relativistic kinetic energy, Lieb and Yau (1988) proved that stability of matter holds in the sense formulated in Theorem 2 if $\alpha (= c^{-1})$ is small enough and $\max_j \{z_j\} \alpha \leq 2/\pi$. It is known here that the value $2/\pi$ is the best possible, since it is so for the one-atom case. The one-atom case had been studied by Herbst. The corresponding case of a one-electron molecule was studied by Lieb and Daubechies. Less optimal results on the stability of matter with relativistic kinetic energy had been

obtained prior to the work of Lieb and Yau by Conlon and later by Fefferman and de la Llave. References to these works can be found in the work of Lieb and Yau (1988).

The relativistic kinetic energy T_j^{Rel} agrees with the free Dirac operator on the positive spectral subspace of the free Dirac operator (i.e., a subspace of $L^2(\mathbb{R}^3; \mathbb{C}^4)$). Therefore, the stability of matter follows if T_j is replaced by the free Dirac operator and if one restricts to the Hilbert space obtained as in [2] but with $L^2(\mathbb{R}^3; \mathbb{C}^2)$ replaced by the positive spectral subspace of the free Dirac operator. This formulation is often referred to as the “no-pair” model. In the usual Dirac picture, the negative spectral subspace, the Dirac sea, is occupied. As long as one ignores pair creation, only the positive spectral subspace is available.

Magnetic fields may be included by considering the “magnetic kinetic energy”

$$T_j^{\text{Mag}} = \frac{1}{2}(-i\nabla_j - c^{-1}A(x_j))^2$$

It turns out that the stability of matter theorem (Theorem 2) holds for all magnetic vector potentials $A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ with a constant $C_{|z|}$ independent of A . This is, therefore, also the case if we consider the magnetic field (or rather the vector potential) as a dynamic variable and add the (positive) field energy

$$\mathcal{U} = \frac{1}{8\pi} \int_{\mathbb{R}^3} |\nabla \times A(x)|^2 dx \quad [4]$$

to the Hamiltonian. The resulting Hamiltonian describes a charged spinless particle interacting with a classical electromagnetic field.

A more complicated situation is described by the “magnetic Pauli kinetic energy”

$$T_j^{\text{Pauli}} = \frac{1}{2}((-i\nabla_j - c^{-1}A(x_j)) \cdot \sigma_j)^2$$

where the coupling of the spin to the magnetic field is included through the vector of 2×2 Pauli matrices acting on the spin components of particle j , that is, $\sigma = (\sigma_1, \sigma_2, \sigma_3)$, with

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

For the Pauli kinetic energy, stability of matter will not hold independently of the magnetic field (or even for a fixed unbounded magnetic field) unless the field energy \mathcal{U} in [4] is included in the Hamiltonian. If the field energy is included, stability of matter holds independently of the magnetic field, that is, even if one minimizes over the dynamic variable A , if $\alpha (= c^{-1})$ and $\max_j \{z_j\} \alpha^2$ are small enough. This was proved by Fefferman (1997) and by Lieb *et al.*

(1995). The latter result includes the physical value of α . The fact that a bound on α is needed had been proved by Loss and Yau. Stability for a one-electron atom had been proved in this model by Fröhlich, Lieb, and Loss. The many-electron atom and the one-electron molecule had been studied by Lieb and Loss. Most relevant references may be found in the work of Lieb *et al.* (1995).

The possibility of quantizing the magnetic field has also been studied. In this case, one must introduce an ultraviolet cutoff in the momentum modes of the vector potential. Stability of matter in the resulting model of (ultraviolet cutoff) QED coupled to non-relativistic matter was proved by Fefferman *et al.* improving results of Bugliaro, Fröhlich, and Graf.

Finally, one may include both relativistic effects and electromagnetic interactions. Let us first discuss the case of classical electromagnetic fields. If instead of the Pauli kinetic energy one uses the Dirac operator with a magnetic vector potential then there would be no lower bound on the energy. But, as previously described, one can study a no-pair formulation of relativistic particles coupled to electromagnetic fields. The question arises which subspace of $L^2(\mathbb{R}^3; \mathbb{C}^4)$ one should restrict to (i.e., which subspace is filled and which one is available). There are two obvious choices. Either one should, as before, restrict to the positive spectral subspace of the free Dirac operator or one should restrict to the positive spectral subspace of the magnetic Dirac operator. It is proved by Lieb *et al.* (1997) that the former choice leads to instability, whereas stability of matter holds for the latter choice under some conditions on α and $\max_j \{z_j\}$. Stability requires that the field energy \mathcal{U} is included in the Hamiltonian. It then holds independently of the magnetic field.

This final stability result also holds if the magnetic field is quantized with an ultraviolet cutoff as proved by Lieb and Loss (2002).

The no-pair model even with the ultraviolet cutoff quantized field is not fully relativistically invariant. As mentioned above, there is still no mathematical formulation of QED, a fully relativistically invariant model for quantum particles interacting with electromagnetic fields.

The Proof of Stability of the First Kind

The proof of stability of the first kind will now be sketched for charged quantum systems.

As mentioned in the introduction, stability of the first kind is a consequence of the uncertainty principle. Contrary to what is often stated in physics textbooks, stability does not follow from the Heisenberg formulation of the uncertainty principle.

A mathematically more flexible formulation is provided by the classical Sobolev inequality, which states that for all square-integrable functions $\psi \in L^2(\mathbb{R}^3)$, one has

$$\int |\nabla \psi|^2 \geq C_S \left(\int |\psi|^6 \right)^{1/3} \quad [5]$$

for $C_S > 0$. It follows from this inequality that for any attractive potential V , there is a lower bound on the energy expectation

$$\begin{aligned} & \left(\psi, \left(-\frac{1}{2} \Delta - V \right) \psi \right) \\ &= \frac{1}{2} \int |\nabla \psi|^2 - \int V |\psi|^2 \geq \frac{1}{2} C_S \left(\int \psi^6 \right)^{1/3} \\ & \quad - \left(\int V^{5/2} \int |\psi|^2 \right)^{2/5} \left(\int |\psi|^6 \right)^{1/5} \\ & \geq -C \int V^{5/2} \int |\psi|^2 \end{aligned}$$

for some $C > 0$. Thus, the lowest possible energy of one particle moving in the potential V is bounded below by $-C \int V^{5/2}$. For N (noninteracting) particles, the lower bound is $-CN \int V^{5/2}$. This holds whether or not the particles have spin. If, more generally, the potential can be written as $V = U + W$, $U, W \geq 0$, where $\int U^{5/2} < \infty$ and W is bounded $W \leq \|W\|_\infty$, then the energy of N noninteracting particles moving in the potential V is bounded below by

$$-NC \int U^{5/2} - N\|W\|_\infty \quad [6]$$

For the Hamiltonian H_N from [1], one can get a lower bound on the energy $E(z, N, K)$ by ignoring all the positive potential terms, that is, the last two sums in [1]. The remaining Hamiltonian describes N independent particles moving in the potential

$$-V = -\sum_{k=1}^K \frac{z_k}{|x - r_k|} = -\sum_{k=1}^K (U_k + W_k)$$

where U_k is the restriction of $z_k/|x - r_k|$ to the set $|x - r_k| < R$ for some $R > 0$ and W_k is the restriction to the complementary set. Using [6], one can easily see that the energy expectation is bounded below by

$$\begin{aligned} & -CNK^{5/2} \max_k \{z_k\}^{5/2} R^{1/2} - NK \max_k \{z_k\} R^{-1} \\ &= -C' NK^2 \max_k \{z_k\}^2 \end{aligned}$$

where we have made the optimal choice for $R \sim (K \max_k \{z_k\})^{-1}$.

This finite lower bound on the energy proves the stability of the first kind, but it clearly does

not have the form required for the stability of the second kind.

The Proof of Stability of Matter

The proof of stability of the first kind presented in the previous section must be improved in two ways in order to conclude the stability of matter.

For fermions, it turns out that the lower bound in [6] can be improved in such a way that there is no factor N in the first term. This is the content of the bound of Lieb and Thirring discussed in the introduction.

Theorem 5 (Lieb–Thirring inequality 1975). *The sum of all the negative eigenvalues of the operator $-(1/2)\Delta - V(x)$ is bounded below by*

$$-L_{LT} \int V^{5/2}$$

for some constant $L_{LT} > 0$

For N noninteracting fermions moving in the potential V , the lowest possible energy is given by the sum of the N lowest eigenvalues of the operator in the above theorem. Thus, the theorem gives a lower bound on this energy independently of N .

The second point where the argument from the previous section has to be improved is the control of the electrostatic energy. In the above discussion, all repulsive terms have simply been ignored. For stability of matter, a much more delicate bound is needed. Many versions of such bounds have been given going back to the work of Onsager (1939). Here, a result of Baxter (1980) will be used.

Theorem 6 (Baxter's correlation estimate). *For all positions $x_1, \dots, x_N, r_1, \dots, r_K \in \mathbb{R}^3$ and all charges $z_1, \dots, z_K > 0$, we have the pointwise inequality*

$$\begin{aligned} & -\sum_{k=1}^K \sum_{i=1}^N \frac{z_k}{|x_i - r_k|} + \sum_{1 \leq i < j \leq N} \frac{1}{|x_i - x_j|} \\ & + \sum_{1 \leq k < l \leq K} \frac{z_k z_l}{|r_k - r_l|} \geq -\sum_{i=1}^N V(x_i) \end{aligned}$$

where $V(x) = (1 + 2 \max_k \{z_k\}) \max_k \{|x - r_k|^{-1}\}$.

This theorem simply states that, for a lower bound, one can replace the full electrostatic Coulomb energy by the energy of independent electrons moving in the potential where they always see only the closest nuclei (with a modified charge). Baxter (1980) used probabilistic techniques to prove the inequality. An improved version of the inequality was given by Lieb and Yau (1988), with an analytic proof.

Similarly to the argument in the previous section, one can write $V(x) = U(x) + W(x)$, where U is the restriction of V to the set where $\min_k \{|x - r_k|\} < R$ for some $R > 0$ and W is the restriction to the complementary set. It then follows from Baxter's correlation estimate and the Lieb–Thirring inequality that the lowest eigenvalue of the Hamiltonian H_N on the fermionic Hilbert space \mathcal{H}_N^F is bounded below by

$$\begin{aligned} & -L_{\text{LT}} \int U^{5/2} - N(1 + 2 \max_k \{z_k\}) R^{-1} \\ & \geq -C(1 + 2 \max_k \{z_k\})^{5/2} K R^{1/2} \\ & \quad - N(1 + 2 \max_k \{z_k\}) R^{-1} \\ & = -C'(1 + 2 \max_k \{z_k\})^2 (N + K) \end{aligned}$$

where $R \sim (1 + 2 \max_k \{z_k\})^{-1}$. This lower bound is linear in the total particle number $N + K$, as required by stability of matter.

From Thomas–Fermi Theory to Stability of Matter

In this final section, the proof of stability of matter by Lieb and Thirring (1975), where they use the Thomas–Fermi theory, is discussed briefly. First note that there is a dual formulation of the Lieb–Thirring inequality theorem (Theorem 5), which makes the connection to the Sobolev inequality [5] much more transparent.

Theorem 7 (Lieb–Thirring inequality as a kinetic energy bound). *For any normalized antisymmetric (fermionic) wave function $\Psi \in \mathcal{H}_N^F$ we have with $C_{\text{LT}} = \frac{3}{5}(\frac{2}{5}L_{\text{LT}}^{-1})^{2/3}$ the following lower bound on the kinetic energy:*

$$\begin{aligned} & \sum_{i=1}^N \frac{1}{2} \int_{\mathbb{R}^{3N}} \|\nabla_i \Psi(x_1, \dots, x_N)\|^2 dx_1 \cdots dx_N \\ & \geq C_{\text{LT}} \int_{\mathbb{R}^3} \rho(x)^{5/3} dx \end{aligned}$$

where $\|\cdot\|$ is the norm in spin space (\mathbb{C}^{2^N}) and the one-electron density is given by

$$\rho(x) = N \int_{\mathbb{R}^{3(N-1)}} \|\Psi(x, x_2, \dots, x_N)\|^2 dx_2 \cdots dx_N$$

This estimate follows immediately from Theorem 5, which implies that

$$\sum_{i=1}^N \frac{1}{2} \int \|\nabla_i \Psi\|^2 - \int \rho V \geq -L_{\text{LT}} \int V^{5/2}$$

To arrive at Theorem 7, simply choose $V = ((2/5)(L_{\text{LT}}^{-1}\rho)^{2/3})$.

One should compare the Lieb–Thirring kinetic energy bound with the expression $(3/10)(3\pi^2)^{2/3}\rho^{5/3}$ for the (thermodynamic) energy density of a free Fermi gas. One of the yet unproven conjectures is that the Lieb–Thirring bound holds with C_{LT} replaced by the free Fermi constant $(3/10)(3\pi^2)^{2/3}$.

The idea in the Lieb–Thirring proof of stability of matter is to bound the energy below by an expression depending only on the one-electron density. Theorem 7 achieves this for the kinetic energy. What is missing is a lower bound on the electrostatic Coulomb energy depending only on the density. One can show (see Lieb (1976) or Lieb and Thirring (1975)) that, except for an error of the form “ $-\text{const} \times N$,” the total energy expectation $(\Psi, H_N \Psi)$ may be bounded below by

$$\begin{aligned} & C_{\text{LT}} \int \rho^{5/3} - \sum_{k=1}^K \int \rho(x) \frac{z_k}{|x - r_k|} dx \\ & + \frac{1}{2} \iint \frac{\rho(x)\rho(y)}{|x - y|} dx dy + \sum_{1 \leq k < \ell \leq K} \frac{z_k z_\ell}{|r_k - r_\ell|} \quad [7] \end{aligned}$$

Here, as before, ρ is the one-electron density of the N -body wave function Ψ . The expression [7] is the famous Thomas–Fermi energy functional. It has been studied rigorously by Lieb and Simon (1977). The Thomas–Fermi energy is the infimum of the expression (7) over all ρ with $\int \rho = N$. One of the important results about the Thomas–Fermi energy is Teller's no-binding theorem (Lieb and Simon 1977). It states that in Thomas–Fermi theory atoms do not bind to form molecules. This means that the Thomas–Fermi energy is greater than the sum of the individual atomic energies (these energies in turn depend only on the nuclear charges).

The above Thomas–Fermi lower bound on the energy expectation $(\Psi, H_N \Psi)$ together with the no-binding theorem implies stability of matter.

The generalizations to stability of matter discussed earlier are proved in a way similar to the proof presented in the previous section.

See also: *h*-Pseudodifferential Operators and Applications; Quantum Statistical Mechanics: Overview; Schrödinger Operators.

Further Reading

- Baxter JR (1980) Inequalities for potentials of particle systems. *Illinois Journal of Mathematics* 24: 645–652.
- Conlon JG, Lieb EH, and Yau H-T (1988) The $N^{7/5}$ law for charged bosons. *Communications in Mathematical Physics* 116: 417–448.
- Dyson FJ (1967) Ground state energy of a finite system of charged particles. *Journal of Mathematical Physics* 8: 1538–1545.

- Dyson FJ and Lenard A (1967) Stability of matter. I. *Journal of Mathematical Physics* 8: 423–434.
- Dyson FJ and Lenard A (1968) Stability of matter. II. *Journal of Mathematical Physics* 9: 698–711.
- Fefferman CL (1997) Stability of matter with magnetic fields. *CRM Proceedings and Lecture Notes* 12: 119–133.
- Fefferman C, Fröhlich J, and Graf GM (1997) Stability of ultraviolet cutoff quantum electrodynamics with non-relativistic matter. *Communications in Mathematical Physics* 190: 309–330.
- Fisher M and Ruelle D (1966) The stability of many-particle systems. *Journal of Mathematical Physics* 7: 260–270.
- Lieb EH and Lebowitz JL (1972) The constitution of matter: existence of thermodynamics for systems composed of electrons and nuclei. *Advances in Mathematics* 9: 316–398.
- Lieb EH and Thirring WE (1975) Bound for the kinetic energy of fermions which proves the stability of matter. *Physical Review Letters* 35: 687–689.
- Lieb EH (1976) The stability of matter. *Reviews of Modern Physics* 48: 553–569.
- Lieb EH and Simon B (1977) Thomas–Fermi theory of atoms, molecules and solids. *Advances in Mathematics* 23: 22–116.
- Lieb EH (1979) The $N^{5/3}$ law for bosons. *Physics Letters A* 70: 71–73.
- Lieb EH and Thirring WE (1984) Gravitational collapse in quantum mechanics with relativistic kinetic energy. *Annals of Physics*, NY 155: 494–512.
- Lieb EH and Yau H-T (1987) The Chandrasekhar theory of stellar collapse as the limit of quantum mechanics. *Communications in Mathematical Physics* 112: 147–174.
- Lieb EH and Yau H-T (1988) The stability and instability of relativistic matter. *Communications in Mathematical Physics* 118: 177–213.
- Lieb EH (1990) The stability of matter: from atoms to stars. 1989 Gibbs Lecture. *Bulletin of the American Mathematical Society* 22: 1–49.
- Lieb EH, Loss M, and Solovej JP (1995) Stability of matter in magnetic fields. *Physical Review Letters* 75: 985–989.
- Lieb EH, Seidertop H, and Solovej JP (1997) Stability and instability of relativistic electrons in classical electromagnetic fields. *Journal of Statistical Physics* 89: 37–59.
- Lieb EH and Loss M (2002) Stability of a model of relativistic quantum electrodynamics. *Communications in Mathematical Physics* 228: 561–588.
- Lieb EH (2004) The stability of matter and quantum electrodynamics, Proceedings of the Heisenberg Symposium, Munich, Dec. 2001. In: Buschhorn G and Wess J (eds.) *Fundamental Physics – Heisenberg and Beyond*, pp. 53–68. Berlin: Springer (arXiv math-ph/0209034).
- Lieb EH and Solovej JP Ground state energy of the two-component charged Bose gas. *Communications in Mathematical Physics* (to appear).
- Onsager L (1939) Electrostatic interaction of molecules. *Journal of Physical Chemistry* 43: 189–196.
- Solovej JP (2004) Upper bounds to the ground state energies of the one- and two-component charged Bose gases. *Preprint*.

Stability of Minkowski Space

S Klainerman, Princeton University, Princeton, NJ, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The Minkowski space, which is the simplest solution of the Einstein field equations in vacuum, that is, in the absence of matter, plays a fundamental role in modern physics as it provides the natural mathematical background of the special theory of relativity. It is most reasonable to ask whether it is stable under small perturbations. In other words, can arbitrary small perturbations of flat initial conditions lead to developments which are radically different, in the large, from the flat Minkowski space? It turns out to be a highly nontrivial problem as the Einstein equations are of a quasilinear hyperbolic character. Typical systems of this type, in three space dimensions, do form singularities in finite time even for small disturbances of their trivial initial data. To avoid finite-time singularities, we must require that sufficiently small perturbations of Minkowski space are geodesically complete. This, however, is not

enough; one should also insist that the corresponding spacetimes become flat along all possible directions, that is, globally asymptotically flat. This is measured by the decay of the curvature tensor to zero. The precise rate of decay is also of interest. One expects that various null-frame components of the curvature tensor decay at different rates along outgoing null hypersurfaces; this goes under the name of “peeling estimates.” It turns out in fact that we cannot prove geodesic completeness without establishing at the same time sufficiently fast rates of decay to flatness corresponding to at least some peeling.

The problem of stability of Minkowski space is intimately related to that of describing the asymptotic properties of the gravitational field at large distances from an isolated, weakly radiating physical system. Precise laws of gravitational radiation can be deduced from the assumption that the spacetime (M, g) under consideration can be conformally compactified by adding a boundary S , called *skry*, to M so that an appropriate conformal rescaling of g can be extended smoothly to the new manifold (\hat{M}, \hat{g}) with boundary. In reality, the compactified spacetime cannot be smooth at the particular point i^0 corresponding to spacelike infinity. A spacetime

(M, g) is called asymptotically simple (AS) if its conformal completion is smooth everywhere except i^0 and every null geodesic intersects \mathcal{S} at precisely two endpoints. The AS assumption allows one to derive precise decay asymptotic for various curvature components of (M, g) along null geodesics which are referred to as strong peeling. The obvious questions raised by this procedure are: do there exist nontrivial AS spacetimes and, if so, do they contain a sufficiently large class of radiating spacetimes including those which appear in all relevant applications?

Clearly, the two problems mentioned above are related but not equivalent. Asymptotically simple spacetimes verify strong peeling, in particular they are globally asymptotically flat, that is, their curvature tensor tends to zero along all geodesics. Yet, it is perfectly possible that arbitrarily small perturbations of the Minkowski space are geodesically complete and globally asymptotically flat without being asymptotically simple.

The first global stability result of the Minkowski metric was proved by Christodoulou and Klainerman (1993). Their result proves sufficiently strong peeling estimates to allow one to derive the most important properties of gravitational radiation, such as the Bondi mass-law formula, but not as strong as those consistent with asymptotic simplicity. A companion result was proved by Klainerman and Nicolò (2003). Recently, Rodnianski and Lindblad (submitted) have obtained a surprising global stability of Minkowski result for the Einstein vacuum equations in the Lorentz gauge, which provides considerable weaker peeling than Christodoulou and Klainerman (1993) and Klainerman and Nicolò (1999) but is much easier to prove.

The goal of this article is to describe various results obtained since the early 1980s concerning both aspects of the problem of stability of Minkowski mentioned above.

Initial Data Formulation

The proper mathematical context for the stability of Minkowski is that provided by the initial-value problem for vacuum solutions to the Einstein field equations, that is, Ricci flat spacetimes (M, g) , $R_{\mu\nu} = 0$. We recall the following simple definitions:

Definition 1 An initial data set is a triplet (Σ, g, k) consisting of a three-dimensional complete Riemannian manifold (Σ, g) and a 2-covariant symmetric tensor k on Σ satisfying the constraint equations:

$$\nabla^i k_{ij} - \nabla_i \text{tr}_g k = 0, \quad R - |k|^2 + (\text{tr } k)^2 = 0$$

where ∇ is the covariant derivative, R the scalar curvature of (Σ, g) . An initial data set is said to be maximal if $\text{tr}_g k = 0$. This is a gauge condition which can be imposed without loss of generality. For simplicity we shall assume, throughout this article, that all initial data sets we consider are maximal.

Definition 2 An initial data set is said to be flat, or trivial, if it corresponds to a complete spacelike hypersurface in Minkowski space with its induced metric and second fundamental form. An initial data set is said to be asymptotically flat if there exists a system of coordinates (x^1, x^2, x^3) defined in a neighborhood of infinity on Σ , with $r = \sqrt{(x^1)^2 + (x^2)^2 + (x^3)^2}$, relative to which the metric g approaches the Euclidian metric and k approaches zero as $r \rightarrow \infty$. We assume, for simplicity, that Σ has only one end. A neighborhood of infinity means the complement of a sufficiently large compact set on Σ .

Remark 1 Because of the constraint equations, the asymptotic behavior cannot be arbitrarily prescribed. A precise definition of asymptotic flatness has to involve the ADM mass of (Σ, g) . Taking the mass into account, we write

$$g_{ij} = \left(1 + \frac{2M}{r}\right) \delta_{ij} + o(r^{-1})$$

According to the positive-mass theorem, $M \geq 0$ and $M = 0$ implies that the initial data set is flat.

Definition 3 We say that an initial data set is strongly asymptotically flat if, for some $\delta > 1/2$, relative to the coordinate system mentioned above,

$$g_{ij} - \left(1 + \frac{2M}{r}\right) \delta_{ij} = O(r^{-1-\delta}), \quad k_{ij} = O(r^{-2-\delta})$$

as $r \rightarrow \infty$

Moreover, every derivative of $g - (1 + 2M/r)\delta$ and k improves the asymptotics by one.

Definition 4 A Cauchy development of an initial data set (Σ, g, k) is a spacetime manifold (M, g) satisfying the Einstein equations together with an embedding $i: \Sigma \rightarrow M$ such that $i_*(g), i_*(k)$ are the first and second fundamental forms of $i(\Sigma)$ in M . A development is required to be also globally hyperbolic (which means that $i(\Sigma)$ is a Cauchy hypersurface, i.e., each causal curve in M intersects $i(\Sigma)$ at precisely one point) in order to assure the unique dependence of solutions on the data. A future development of (Σ, g, k) consists of a globally hyperbolic manifold (M, g) with boundary, satisfying the Einstein equations, and an embedding i as before which identifies Σ to the boundary of M .

The most primitive question asked about the initial-value problem, solved in a satisfactory way, for very large classes of evolution equations, is that of local existence and uniqueness of solutions. For the Einstein equations, this type of result was first established by Bruhat (1952) with the help of wave coordinates which allowed her to cast the Einstein equations in the form of a system of nonlinear wave equations to which one can apply the standard theory of symmetric hyperbolic systems. A stronger result, due to Hughes *et al.* (1976), states the following:

Theorem 1 *Let (Σ, g, k) be an initial data set for the Einstein vacuum equations. Assume that Σ can be covered by a locally finite system of coordinate charts U_α related to each other by C^1 diffeomorphisms, such that $(g, k) \in H_{\text{loc}}^s(U_\alpha) \times H_{\text{loc}}^{s-1}(U_\alpha)$ with $s > 5/2$. Then there exists a unique (up to an isometry) globally hyperbolic, Hausdorff, development (M, g) for which Σ is a Cauchy hypersurface.*

In Theorem 1, the uniqueness up to an isometry requires additional regularity, $s > (5/2) + 1$, on the data. One has uniqueness, however, without additional regularity for the reduced Einstein equations system in wave coordinates.

Remark 2 In the case of nonlinear systems of differential equations, the local existence and uniqueness result leads, through a straightforward extension argument, to a global result. The formulation of the same type of result for the Einstein equations is a little more subtle; it was done by Bruhat and Geroch.

Theorem 2 (Bruhat–Geroch). *For each smooth initial data set, there exists a unique maximal future development.*

Thus, any construction, obtained by an evolutionary approach from a specific initial data set, must be necessarily contained in its maximal development. This may be said to solve the problem of global existence and uniqueness in general relativity. This is of course misleading, for equations defined in a fixed background global is a solution which exists for all time. In general relativity, however, we have no such background as the spacetime itself is the unknown. The connection with the classical meaning of a global solution requires a special discussion concerning the proper time of timelike geodesics; all further questions may be said to concern the qualitative properties of the maximal development. The central issue is that of existence and character of singularities. First, we can define a regular maximal development as one which is complete in the sense that all future timelike and null geodesics can be indefinitely extended

relative to their proper time (or affine parameter in the case of null geodesics). If the initial data set is sufficiently far off from the trivial one, the corresponding future development may not be regular. This is the content of the following well-known theorem of Penrose (1979).

Theorem 3 *If the manifold support of an initial data set is noncompact and contains a closed trapped surface, the corresponding maximal development is incomplete.*

Stability of Minkowski Space

At the opposite end of Penrose's trapped-surface condition, the problem of stability of Minkowski space concerns the development of asymptotically flat initial data sets which are sufficiently close to the trivial one. Although it may be reasonable to expect the existence of a sufficiently small neighborhood of the trivial initial data set, in an appropriate topology, such that all corresponding developments are geodesically complete and globally asymptotically flat, such a result was by no means preordained. First, all known explicit asymptotically flat solutions of the Einstein vacuum equations, that is, the Kerr family, are singular. The attempts to construct nonexplicit, dynamic, solutions based on the conformal compactification method, due to Penrose (1962), were obstructed by the irregular behavior of initial data sets at i^0 . (The problem is that the singularity at i^0 could propagate and thus destroy the expected smoothness of \mathcal{S}^2 . This problem has been recently solved by constructing initial data sets which are precisely stationary at spacelike infinity.) Finally, the attempts, using partial differential equation hyperbolic methods, to extend the classical local result of Bruhat ran into the usual difficulties of establishing global in time existence to solutions of quasilinear hyperbolic systems. Indeed, as mentioned above, the wave coordinate gauge allows one to express the Einstein vacuum equations in the form of a system of nonlinear wave equations which does not satisfy Klainerman's null condition (the null condition (Klainerman 1983, 1986) identifies an important class of quasilinear systems of wave equations in four spacetime dimensions for which one can prove global in time existence of small solutions) and thus was sought to lead to formation of singularities. (The conjectured singular behavior of wave coordinates was sought, however, to reflect only the instability of the specific choice of gauge condition and not a true singularity of the equations.) According to Bruhat (personal communication),

Einstein himself had reasons to believe that the Minkowski space may not be stable. The problem of stability of the Minkowski space was first settled by Christodoulou and Klainerman (1990).

Theorem 4 (Global stability of Minkowski). *Any asymptotically flat initial data set which is sufficiently close to the trivial one has a complete maximal future development.*

A related result (Theorem 5) proved recently by Klainerman and Nicolò (2003a), solves the problem of radiation for arbitrary asymptotically flat initial data sets: a proof the result below can also be derived, indirectly, from Christodoulou and Klainerman (1993). The proof of Klainerman and Nicolò (2003a) avoids, however, a great deal of the technical complications of this proof.

Theorem 5 *For any, suitably defined, asymptotically flat initial data set (Σ, g, k) with maximal future development (M, g) , one can find a suitable domain $\Omega_0 \subset \Sigma$ with compact closure in Σ such that the boundary \mathcal{D}_0^+ of its domain of influence $C^+(\Omega_0)$, or causal future of Ω , in M has complete null geodesic generators with respect to the corresponding affine parameters.*

Both the results of Christodoulou–Klainerman and Klainerman–Nicolò prove in fact a lot more than stated above. They provide a wealth of information concerning the behavior of null hypersurfaces as well as the rate at which various components of the Riemann curvature tensor approach zero along time-like and null geodesics. Here are more precise versions for Theorems 4 and 5.

Theorem 4 (Expanded version). *Assume that (Σ, g, k) is maximal and strong asymptotically flat, $g - (1 + 2M/r)\delta = O(r^{-3/2})$, $k = O(r^{-5/2})$ plus an appropriate global smallness assumption. We can construct complete spacetime (M, g) together with a maximal foliation Σ_t given by the level hypersurfaces of a time function t and null foliation C_u , given by the level hypersurfaces of an outgoing optical function u such that relative to an adapted null frame $e_4 = L$, $e_3 = \underline{L}$, and $(e_a)_{a=1,2}$ we have, along the null hypersurfaces C_u the weak peeling decay,*

$$\begin{aligned} \alpha_{ab} &= R(L, e_a, L, e_b) = O(r^{-7/2}) \\ 2\beta_a &= R(L, \underline{L}, L, e_a) = O(r^{-7/2}) \\ 4\rho &= R(L, \underline{L}, L, \underline{L}) = O(r^{-3}) \\ 4\sigma &= R(L, \underline{L}, L, \underline{L}) = O(r^{-3}) \\ 2\beta_a &= R(L, \underline{L}, L, e_a) = O(r^{-2}) \\ \underline{\alpha}_{ab} &= R(\underline{L}, e_a, \underline{L}, e_b) = O(r^{-1}) \end{aligned} \quad [1]$$

as $r \rightarrow \infty$ with $4\pi r^2 = \text{Area}(S_{t,u} = \Sigma_t \cap C_u)$. Also, $\rho - \bar{\rho}, \sigma = O(r^{-7/2})$, with $\bar{\rho}$ the average of ρ over the compact 2-surfaces $S_{t,u} = \Sigma_t \cap C_u$.

Three points are noteworthy. (1) The outgoing optical solution refers to the solution of the Eikonal equation $g^{\alpha\beta} \partial_\alpha u \partial_\beta u = 0$ whose level hypersurfaces C_u intersect Σ_t in expanding wave fronts for increasing t ; (2) the generators L and \underline{L} are given by: $L = -g^{\alpha\beta} \partial_\beta u \partial_\alpha$, the null geodesic generator of C_u ; \underline{L} is then the null conjugate of L , perpendicular to $S_{t,u} = C_u \cap \Sigma_t$; and (3) e_a is an orthonormal frame on $S_{t,u}$.

Theorem 5 (Expanded version). *For any asymptotically flat initial data sets (Σ, g, k) , verifying the same asymptotically flat conditions as in Theorem 4 one can find a suitable domain $\Omega_0 \subset \Sigma$ with compact closure in Σ such that its future domain of influence $C^+(\Omega_0)$ can be foliated by two null foliations; one outgoing $C(u)$ whose leaves are complete towards the future and the second one $\underline{C}(u)$ which is incoming. Let $S(u, u) = C(u) \cap \underline{C}(u)$ denote the compact 2-surfaces of intersection between the outgoing and incoming null hypersurfaces, whose area is denoted by $4\pi r^2$, and consider an adapted null frame (that is, L is a the geodesic null generator of $C(u)$, \underline{L} its null conjugate perpendicular to $S(u, u)$, and e_a an orthonormal frame on $S(u, u)$) $L, \underline{L}, (e_a)_{a=1,2}$ at every point along an outgoing null cone $C(u)$. Then, denoting by $\alpha, \beta, \rho, \sigma, \underline{\beta}, \underline{\alpha}$ the null components of the curvature tensor, as in Theorem 5, we have, along $C(u)$ as $r \rightarrow \infty$,*

$$\begin{aligned} \alpha, \beta, \rho - \bar{\rho}, \sigma &= O(r^{-7/2}), \quad \underline{\beta} = O(r^{-2}), \\ \underline{\alpha} &= O(r^{-1}) \end{aligned} \quad [2]$$

Observe that the rates of decay in [1] and [2] are the same. This will be referred to as weak peeling to distinguish from the rates of decay compatible with asymptotic simplicity, that is,

$$\begin{aligned} \alpha &= O(r^{-5}), \quad \beta = O(r^{-4}) \\ \rho, \sigma &= O(r^{-3}), \quad \underline{\beta} = O(r^{-2}), \quad \underline{\alpha} = O(r^{-1}) \end{aligned} \quad [3]$$

to which we shall refer as strong peeling. We shall discuss more about these in the next section, following a review, of a recent result of Lindblad–Rodnianski.

Even the expanded forms of Theorems 4 and 5 stated here do not exhaust, all the information provided by global stability results in Christodoulou and Klainerman (1993) and Klainerman and Nicolò (2003a). Of particular interest are the main asymptotic conclusions which can be derived with the help of these information, the most

important being the Bondi mass-law formula which calculates the gravitational energy radiated at null infinity.

The simplest gauge condition in which the hyperbolic character of the Einstein field equations are easiest to exhibit is the wave coordinate condition; that is, one solves the Einstein vacuum equations relative to a special system of coordinates x^α which satisfy the equation $\square g x^\alpha = 0$. Then, denoting by $h_{\alpha\beta} = g_{\alpha\beta} - m_{\alpha\beta}$ with m the standard Minkowski metric, we obtain the following system of quasilinear wave equations in h ,

$$g^{\mu\nu} \partial_\mu \partial_\nu h = N(h, \partial h) \quad [4]$$

with $N(h, \partial h)$ a nonlinear term, quadratic in ∂h , which can be exhibited explicitly. This form of the Einstein field equations, called the wave coordinates reduced Einstein equations, is precisely the one which allowed Bruhat (1952) to prove the first local existence result. Later, she also pointed out that the first nontrivial iterate of [4] behaves like $t^{-1} \log t$ rather than t^{-1} as expected from the decay properties of solutions to $\square h = 0$ in Minkowski space. This seems to indicate that the wave coordinates may not be suitable to study the long-time behavior of solutions to the Einstein field equations. This negative conclusion is also consistent with the fact that the eqns [4] do not verify Klainerman's null condition. (Klainerman's null condition (Klainerman 1983) is an algebraic condition on systems of nonlinear wave equations in $(1+3)$ dimensions, similar to [4], which allows one to extend all local solutions, corresponding to small initial data, for all time. Moreover, these solutions decay at the rate of t^{-1} as $t \rightarrow \infty$ consistent to the decay of free waves.) Lindblad and Rodnianski (2003) were able to isolate a new condition, which they call the weak null condition, verified by the wave coordinates reduced Einstein eqns [4], for which one can prove a small data global existence result consistent with the weaker decay rates suggested by the linear asymptotic analysis of Bruhat. Although the new result provides far weaker peeling information than [1], it is much simpler to prove than both Theorems 4 and 5. Moreover, the result seems to apply to a broader class of initial data than in Theorems 4 and 5. It remains an intriguing open problem whether the result of Lindblad–Rodnianski can be used as a stepping stone towards the more complete results of Theorems 4 and 5; that it is once a complete solution, with limited peeling, is known to exist whether one can improve, using the more precise techniques employed in Theorems 4 and 5 minus an

important part of their technical complications, the weak peeling properties of [1].

Strong Peeling

The weak peeling properties [1] derived in Theorems 4 and 5 are consistent, from a scaling point of view, with the SAF condition. To derive strong peeling, see [3], one needs stronger asymptotic conditions. Recently, Corvino–Schoen and Chruściel and Delay (2002) have proved the existence of a large class of asymptotically flat initial data sets (Σ, g, k) which are precisely stationary (here $g_{\text{kerr}}, k_{\text{kerr}}$ are the initial data of the a Kerr solution in standard coordinates) $g = g_{\text{kerr}}, k = k_{\text{kerr}}$ outside a sufficiently large compact set. Moreover, they have proved the existence of sufficiently small solutions in this class which satisfy the requirements needed in Friedrich's conformal compactification method (see Friedrich (2002) and the references within) to produce asymptotically simple spacetimes, that is, spacetimes satisfying Penrose's regular compactification condition (Penrose 1962). Simultaneously, Klainerman and Nicolò (1999) were able to refine the methods used in the proof of Theorem 5 to prove the following:

Theorem 6 *Assume that the initial data set (Σ, g, k) of Theorem 5 satisfies the stronger assumption,*

$$g - g_S = O(r^{-(3/2+\gamma)}), \quad k = (r^{-(5/2+\gamma)}) \quad [5]$$

for some $\gamma > 3/2$. Here

$$g_S = \left(1 - 2\frac{M}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2 \theta d\phi^2)$$

denotes the restriction of the Schwarzschild to $t=0$ in standard polar coordinates. Then, in addition to the results reported in Theorem 5, we have the strong peeling estimates,

$$\alpha = O(r^{-5}), \quad \beta = O(r^{-4})$$

as $r \rightarrow \infty$ along the outgoing null leaves $C(u)$. Moreover, the same conclusions hold true if [5] is replaced by

$$g - g_{\text{kerr}} = O(r^{-(3/2+\gamma)}), \quad k - k_{\text{kerr}} = (r^{-(5/2+\gamma)}) \quad [6]$$

for some $\gamma > 5/2$.

The first part of the theorem was proved in Klainerman and Nicolò (2003b). The second part is work in progress by Klainerman and Nicolò. The existence of initial conditions of the type required in Theorem 6 was established in the works of Corvino (2000) and Chruściel and Delay (2002).

Open Problems

Problem 1 *Extend results of Theorems 5 and 6 to the whole domain of dependence, for small data sets.*

The results of Theorems 5 and 6 give a satisfactory description of gravitational radiation of general classes of asymptotically flat initial data sets outside the domain of dependence of a sufficiently large compact set. It would be desirable to extend these results to the whole domain of dependence of initial data sets which satisfy an additional global smallness assumption similar to that of Theorem 4.

Problem 2 *Is strong peeling (and implicitly asymptotic simplicity) consistent with physically relevant data? If not, is weak peeling a good substitute?*

Damour and Christodoulou (2000) have given conclusive evidence that under no-incoming-radiation condition the future null infinity cannot be smooth. In fact, $\beta = O(r^{-4} \log r)$ as $r \rightarrow \infty$.

Problem 3 *Can one weaken the AF conditions to include, for example, initial data sets with infinite ADM angular momentum?*

It is reasonable to expect a global stability of Minkowski result for small initial data sets which verify, for arbitrarily small ϵ ,

$$g - \left(1 + 2\frac{M}{r}\right)\delta = O(r^{-1-\epsilon}), \quad k = O(r^{-2-\epsilon})$$

One expects in this case that the top null components α and β decay only like $O(r^{-3})$ as $r \rightarrow \infty$ along the null hypersurfaces $C(u)$. It seems that the methods of Lindblad–Rodnianski can treat this case but can only give decay estimates for α, β of the form $O(r^{-3+\epsilon})$.

Problem 4 *Is the Kerr solution in the exterior of the black hole stable?*

The problem remains wide open.

See also: Asymptotic Structure and Conformal Infinity; Classical Groups and Homogeneous Spaces; Critical Phenomena in Gravitational Collapse; Einstein Equations: Exact Solutions; Geometric Analysis and General Relativity; Supergravity.

Further Reading

Bruhat YC (1952) Theoreme d'existence pour certain systemes d'equations aux derivees partielles nonlineaires. *Acta Mathematica* 88: 141–225.

- Bruhat YCh and Geroch RP Global aspects of the Cauchy problem in general relativity. *Communications in Mathematical Physics* 14: 329–335.
- Christodoulou D (2000) *The Global Initial Value Problem in General Relativity*, Lecture Given at the Ninth Marcel Grossman Meeting, 2–8 July, Rome.
- Christodoulou D and Klainerman S (1990) Asymptotic properties of linear field equations in Minkowski space. *Communications in Pure and Applied Mathematics* XLIII: 137–199.
- Christodoulou D and Klainerman S (1993) The global non-linear stability of the Minkowski space. *Princeton Mathematical Series*.
- Chruściel PT and Delay E (2002) Existence of non-trivial, vacuum, asymptotically simple spacetimes. *Classical and Quantum Gravity* 19: L71–L79.
- Corvino J (2000) Scalar curvature deformation and a gluing construction for the Einstein constraint equations. *Communications in Mathematical Physics* 214: 137–189.
- Friedrich H (2002) Conformal Einstein evolution. In: Frauendiener J and Friedrich H (eds.) *The Conformal Structure of Space-Time*, Lecture Notes in Physics. Springer.
- Hughes T, Kato T, and Marsden J (1976) Well-posed quasilinear second-order hyperbolic systems. *Archives for Rational and Mechanical Analysis* 63: 273–294.
- Hawking SW and Ellis GFR (1973) The large scale structure of spacetime. *Cambridge Monographs on Mathematical Physics*.
- Klainerman S (1983) Long-Time Behavior of Solutions to Non-linear Wave Equations Proc. ICM Warszawa.
- Klainerman S (1986) The null condition and global existence to non-linear wave equations. *Lectures in Applied Mathematics* 23: 293–326.
- Klainerman S and Nicolò F (1999) On local and global aspects of the Cauchy problem in general relativity. *Classical and Quantum Gravity* 16: R73–R157.
- Klainerman S and Nicolò F (2003a) The evolution problem in general relativity. *Progress in Mathematical Physics*, vol. 25. Boston: Birkhäuser.
- Klainerman S and Nicolò F (2003b) Peeling properties of asymptotically flat solutions to the Einstein vacuum equations. *Classical and Quantum Gravity*.
- Lindblad H and Rodnianski I (2003) The weak null condition for Einstein's equations. *Comptes Rendus Hebdomadaires des Seances de l'Academie des Sciences, Paris* 336(11): 901–906.
- Lindblad H, and Rodnianski I Global existence for the Einstein vacuum equations in wave coordinates. *Annals of Mathematics* (submitted).
- Penrose R (1965) Zero rest-mass fields including gravitation: asymptotic behaviour. *Proceedings of the Royal Society of London. Series A* 284: 159–203.
- Penrose R (1979) Singularities and time asymmetry in general relativity – an Einstein centenary survey. In: Hawking S and Israel W (eds.) Cambridge: Cambridge University Press.
- Penrose R (1962) Zero rest mass fields including gravitation: asymptotic behavior. *Proceedings of the Royal Society of London. Series A* 270: 159–203.
- Wald R (1984) *General Relativity*. University of Chicago Press.

Stability Problems in Celestial Mechanics

A Celletti, Università di Roma "Tor Vergata," Rome, Italy

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The long-term stability of planets and satellites might be desumed by the regular dynamics that we constantly observe. However, the ultimate fate of the solar system is an intriguing question, which has puzzled scientists since antiquity. In the past centuries, the common belief of a regular motion of the main planets was strengthened by the discovery of a simple law, due to J D Titius and J E Bode (eighteenth century), which provides a recipe to compute the approximate distances of the planets from the Sun. Adopting astronomical units as a measure of the distance, the Titius–Bode law can be stated as

d_n = 0.4 + 0.3 \times 2^n AU [1]

where the index *n* must be selected as provided in Table 1, which compares the distances computed according to [1] with the observed values. Titius and Bode already noticed that it was necessary to skip one unit in *n* from Mars to Jupiter; indeed, the quantity *d*₃ = 2.8 AU might correspond to an average distance of some minor bodies of the asteroid belt, which had been discovered since the beginning of the nineteenth century. The studies of the *N*-body problem, namely the dynamics of *N* mutually attracting bodies (according to Newton’s law), inspired several mathematical and physical theories: from the development of perturbation methods to the discovery of chaotic systems, as attested by the masterly work of H Poincaré (1892). In particular, perturbation theory had relevant applications in celestial mechanics; for example, it led to the prediction of the existence of Neptune in the nineteenth century by J C Adams and U Leverrier

and later to the discovery of Pluto by C Tombaugh, as a result of unexplained perturbations on Uranus and Neptune, respectively. Modern advances in perturbation theories have been provided by the Kolgomorov–Arnol’d–Moser (KAM) and Nekhoroshev theorems, which find broad applications in celestial mechanics insofar as simple model problems are concerned.

The stability of the solar system can also be approached through numerical investigations, which allow one to predict the motion of the celestial bodies using more realistic models. The results of the numerical integrations undermine in some cases the apparent regularity of the solar system: in the following sections, we shall review many examples of regular and chaotic motions in different contexts of celestial mechanics, from the *N*-body problem to the rotational dynamics.

The Restricted Three-Body Problem

Let *P*₁, ..., *P*_{*N*} be *N* bodies with masses *m*₁, ..., *m*_{*N*}, which interact through Newton’s law. Let *u*^(*i*) ∈ *R*³, *i* = 1, 2, ..., *N*, denote the position of the bodies in an inertial reference frame. Normalizing the gravitational constant to 1, the equations of motion of the *N*-body problem have the form

d^2u^{(i)}/dt^2 = - \sum_{j=1, j \neq i}^N \frac{m_j(u^{(i)} - u^{(j)})}{|u^{(i)} - u^{(j)}|^3}, \quad i = 1, \dots, N [2]

In the case *N* = 2, one reduces to the two-body problem, which can be explicitly solved by means of Kepler’s laws as follows. Consider, for example, the Earth–Sun case: for negative values of the energy, the trajectory of the Earth is an ellipse with one focus coinciding with the barycenter, which can practically be identified with the Sun; the Earth–Sun radius vector describes equal areas in equal times; the cube of the semimajor axis is proportional to the square of the period of revolution.

Consider now an extension to the study of three bodies such that in the Keplerian approximation *P*₂ and *P*₃ move around *P*₁ and such that the semimajor axis of *P*₂ is greater than that of *P*₃ (an example is obtained identifying *P*₁ with the Sun, *P*₂ with the Jupiter, and *P*₃ with an asteroid of the main belt). The three-body problem is described by [2] setting *N* = 3; a special case is given by the restricted three-body problem, which describes the evolution of a “zero-mass” body under the gravitational attraction exerted by an assigned two-body system. Setting *N* = 3 and *m*₃ = 0 in [2], the

Table 1 Tititus–Bode law and observed data

	Index <i>n</i> (of [1])	Distance computed from [1]	Observed distance (AU)
Mercury	−∞	0.4	0.39
Venus	0	0.7	0.72
Earth	1	1	1
Mars	2	1.6	1.52
Jupiter	4	5.2	5.2
Saturn	5	10	9.54
Uranus	6	19.6	19.19

equations governing the restricted three-body problem are given by

$$\begin{aligned}\frac{d^2 u^{(1)}}{dt^2} &= -\frac{m_2(u^{(1)} - u^{(2)})}{|u^{(1)} - u^{(2)}|^3} \\ \frac{d^2 u^{(2)}}{dt^2} &= -\frac{m_1(u^{(2)} - u^{(1)})}{|u^{(2)} - u^{(1)}|^3} \\ \frac{d^2 u^{(3)}}{dt^2} &= -\frac{m_1(u^{(3)} - u^{(1)})}{|u^{(3)} - u^{(1)}|^3} - \frac{m_2(u^{(3)} - u^{(2)})}{|u^{(3)} - u^{(2)}|^3}\end{aligned}$$

The first two equations concern the motion of the primaries P_1 and P_2 and they correspond to a Keplerian two-body problem, whose solution can be inserted in the equation for $u^{(3)}$, which becomes a periodically forced second-order equation. The restricted three-body problem can be conveniently described in terms of suitable action-angle coordinates, known as Delaunay variables. The present discussion is restricted to the planar case, namely we assume that the motion of the three bodies takes place on the same plane. The corresponding Delaunay variables, say $(L, G, \ell, \gamma) \in \mathbb{R}^2 \times \mathbb{T}^2$, are defined as follows (Szebehely 1967). Let a and e be, respectively, the semimajor axis and the eccentricity of the osculating orbit of P_3 and let $\mu = 1/m_1^{2/3}$; then Delaunay's action variables are given by

$$L = \mu\sqrt{m_1 a}, \quad G = L\sqrt{1 - e^2}$$

Next, introduce the angle variables: we denote by λ and φ the longitudes of Jupiter and of the asteroid; let γ be the argument of perihelion, namely the angle formed by the periapsis direction with a preassigned reference line, and let u denote the eccentric anomaly, which can be defined through

$$\tan \frac{\varphi - \gamma}{2} = \sqrt{\frac{1+e}{1-e}} \tan \frac{u}{2} \quad [3]$$

Let ℓ be the mean anomaly, which is related to the eccentric anomaly by means of Kepler's equation

$$\ell = u - e \sin u \quad [4]$$

Delaunay's angle variables are represented by the mean anomaly ℓ and by the argument of perihelion γ . For completeness, it should be remarked that the distance r between the minor body P_3 and the primary P_1 is related to the longitude and to the eccentric anomaly by means of the relations

$$r = \frac{a(1 - e^2)}{1 + e \cos(\varphi - \gamma)} = a(1 - e \cos u) \quad [5]$$

In a reference frame centered at one of the primaries, say P_1 , let $H = H(L, G, \ell, \gamma, \lambda)$ denote the Hamiltonian function describing the planar

problem; notice that $H(L, G, \ell, \gamma, \lambda)$ has two degrees of freedom and an explicit time dependence through the longitude λ of P_2 . If the primaries are assumed to move in circular orbits around their common center of mass, the Hamiltonian function reduces to two degrees of freedom, where a new variable g is introduced as the difference between the argument of perihelion γ and the longitude λ of the primary. Normalizing the units of measure so that the distance between the primaries and the sum of their masses is unity, the Hamiltonian function H describing the circular, planar, restricted three-body problem is given by

$$H(L, G, \ell, g) = -\frac{1}{2L^2} - G + \varepsilon F(L, G, \ell, g) \quad [6]$$

where $\varepsilon = \mu m_2$. The perturbing function takes the form

$$F = r \cos(f + g) - \frac{1}{\sqrt{1 + r^2 - 2r \cos(f + g)}}$$

where $f = \varphi - \gamma$ represents the true anomaly, namely the angle formed by the instantaneous orbital radius with the periapsis line. Notice that the quantities r and f are functions of the Delaunay variables through the relations [3]–[5]. As a consequence, one can expand the perturbing function in the form (Delaunay 1860)

$$F(L, G, \ell, g) = \sum_{j,k \geq 0} F_{jk}(\ell, g) e^j a^k$$

where F_{jk} are cosine terms with arguments given by a linear combination of the variables ℓ and g . For example, the first few terms of the series development are given by the following expression:

$$\begin{aligned}F(L, G, \ell, g) &= -1 - \frac{L^4}{4} - \frac{9}{64} L^8 + \frac{L^4 e}{2} \cos \ell \\ &\quad - \left(\frac{3}{8} L^6 + \frac{15}{64} L^{10} \right) \cos(\ell + g) \\ &\quad + \frac{9}{4} L^4 e \cos(\ell + 2g) \\ &\quad - \left(\frac{3}{4} L^4 + \frac{5}{16} L^8 \right) \cos(2\ell + 2g) \\ &\quad - \frac{3}{4} L^4 e \cos(3\ell + 2g) \\ &\quad - \left(\frac{5}{8} L^6 + \frac{35}{128} L^{10} \right) \cos(3\ell + 3g) \\ &\quad - \frac{35}{64} L^8 \cos(4\ell + 4g) \\ &\quad - \frac{63}{128} L^{10} \cos(5\ell + 5g) + \dots \quad [7]\end{aligned}$$

where the eccentricity is a function of the actions through $e = \sqrt{1 - G^2/L^2}$. We remark that the Hamiltonian [6] is nearly integrable with perturbing parameter ε ; indeed, for $\varepsilon = 0$ one recovers the two-body problem describing the interaction between P_1 and P_3 , which can be explicitly solved according to Kepler's laws.

KAM Stability

Classical perturbation theory, as developed by Laplace, Lagrange, Delaunay, Poincaré, etc., does not allow investigation of the stability of the N -body problem, since the series defining the solution are generally divergent. In order to justify this statement, let us start by rewriting the unperturbed Hamiltonian in [6] as

$$h(L, G) = -\frac{1}{2L^2} - G \quad [8]$$

so that [6] becomes $H(L, G, \ell, g) = h(L, G) + \varepsilon F(L, G, \ell, g)$. In order to remove the perturbation to the second order in the perturbing parameter, one looks for a change of variables $(L, G, \ell, g) \rightarrow (L', G', \ell', g')$ close to the identity, that is,

$$\begin{aligned} L &= L' + \varepsilon \frac{\partial \Phi}{\partial \ell}(L', G', \ell, g) \\ G &= G' + \varepsilon \frac{\partial \Phi}{\partial g}(L', G', \ell, g) \\ \ell' &= \ell + \varepsilon \frac{\partial \Phi}{\partial L'}(L', G', \ell, g) \\ g' &= g + \varepsilon \frac{\partial \Phi}{\partial G'}(L', G', \ell, g) \end{aligned}$$

where $\Phi(L', G', \ell, g)$ is the generating function of the transformation. Let

$$\frac{\partial h}{\partial L}(L, G) = \frac{1}{L^3} \equiv \omega(L)$$

In order to perform a first-order perturbation theory, we look for a generating function $\Phi(L', G', \ell, g)$, such that the transformed Hamiltonian is integrable up to $O(\varepsilon^2)$, namely

$$\begin{aligned} &h\left(L' + \varepsilon \frac{\partial \Phi}{\partial \ell}(L', G', \ell, g), G' + \varepsilon \frac{\partial \Phi}{\partial g}(L', G', \ell, g)\right) \\ &+ \varepsilon F\left(L' + \varepsilon \frac{\partial \Phi}{\partial \ell}(L', G', \ell, g), G' + \varepsilon \frac{\partial \Phi}{\partial g}(L', G', \ell, g), \ell, g\right) \\ &= h_1(L', G') + \varepsilon \left[\omega(L') \frac{\partial \Phi}{\partial \ell}(L', G', \ell', g') \right. \\ &\quad \left. - \frac{\partial \Phi}{\partial g}(L', G', \ell', g') + F(L', G', \ell', g') \right] + O(\varepsilon^2) \end{aligned}$$

where $h_1(L', G')$ is the new unperturbed Hamiltonian. If we denote by $F_0(L', G')$ the average of the perturbing function over the angle variables, the new unperturbed Hamiltonian takes the form

$$h_1(L', G') = h(L', G') + \varepsilon F_0(L', G')$$

Expanding F in Fourier series as $F(L, G, \ell, g) = \sum_{n, m \in \mathbb{Z}} F_{nm}(L, G) e^{i(n\ell + mg)}$, the generating function is given by the following expression:

$$\Phi(L', G', \ell, g) = -i \sum_{n, m \in \mathbb{Z} \setminus \{0\}} \frac{F_{nm}(L', G')}{\omega(L')n - m} e^{i(n\ell + mg)}$$

The occurrence of small divisors of the form

$$\frac{1}{\omega(L)n - m}, \quad n, m \in \mathbb{Z}$$

might prevent the convergence of the series defining the generating function. In particular, we remark that zero divisors occur whenever $\omega(L) = m/n$. This situation, which is called an $m:n$ orbit-orbit resonance, implies that during a given interval of time the body P_3 makes m revolutions, whereas P_2 makes exactly n orbits about P_1 .

The control of the occurrence of the small divisors was obtained through a theorem by A N Kolmogorov, who made a major breakthrough in the study of nearly integrable systems. He proved, under general assumptions, that some regions of the phase space are almost filled by maximal invariant tori. The theorem provides a constructive algorithm to give estimates on the perturbing parameter, ensuring the existence of some invariant surfaces. Kolmogorov's theorem was later extended by V I Arnol'd and J Moser, giving rise to the so-called KAM theory. More precisely, the KAM theorem can be stated as follows (see, e.g., Arnol'd *et al.* (1997)): consider a real-analytic, nearly integrable Hamiltonian function and fix a rationally independent frequency vector ω ; if the unperturbed Hamiltonian is not degenerate and if the frequency satisfies a strong nonresonance assumption (called the diophantine condition), for sufficiently small values of the perturbing parameter, there exists an invariant torus on which a quasiperiodic motion with frequency ω takes place. A preliminary investigation of the stability of the N -body problem by means of KAM theory (Arnol'd *et al.* 1997) leads to the existence of large regions filled by quasiperiodic motions, provided the masses of the planets are sufficiently small. Arnol'd's version of KAM theorem has been applied by J Laskar and P Robutel to the spatial three-body planetary problem (the planetary problem concerns the study of the

dynamics of two bodies with comparable masses, moving in the gravitational field of a larger primary) and the existence of quasiperiodic motions has been proved for values of the ratio of semimajor axis less than 0.8 and for inclinations up to $\sim 1^\circ$.

Concrete estimates on the strength of the perturbation were given by M Hénon: in the context of the three-body problem, the application of the original version of Arnol'd's theorem allows one to prove the existence of invariant tori for values of the perturbing parameter (representing the Jupiter–Sun mass ratio) $\leq 10^{-333}$ while the implementation of Moser's theorem provides an estimate of 10^{-50} . We remark that the astronomical value of the Jupiter–Sun mass ratio amounts to $\sim 10^{-3}$, showing a relevant discrepancy between KAM results and physical measurements. More recently, KAM estimates have been refined and adapted to the study of significant problems of celestial mechanics (Celletti and Chierchia 1995). Strong improvements have been obtained combining accurate estimates with a computer-assisted implementation, where the computer is used to perform long computations concerning the development of the perturbing series and the check of KAM estimates. The numerical errors are controlled through the implementation of a suitable technique, known as interval arithmetic. In the framework of the planar, circular, restricted three-body problem, the stability of some asteroids has been proved by A Celletti and L Chierchia for realistic values of the perturbing parameter (e.g., for $\varepsilon = 10^{-3}$). A suitable approximation of the disturbing function (namely, a finite truncation of the series development as in [7]) has been considered. The result relies on an implementation of a computer-assisted isoenergetic KAM theorem and on the following remark: in the four-dimensional phase space, on a fixed energy level the invariant two-dimensional surfaces separate the phase space, providing the stability of the actions for all motions trapped between any two invariant tori. Since the action variables are related to the semimajor axis and to the eccentricity of the orbit, one obtains that the elliptic elements remain close to their initial values.

A computer-assisted KAM theorem has been applied by A Giorgilli and U Locatelli to the planetary (Jupiter–Saturn) problem. Using a suitable secular approximation, it can be shown that this model admits two invariant tori, which bound the orbits corresponding to the initial data of Jupiter and Saturn.

Nekhoroshev Stability

A different approach in order to study the stability of nearly integrable systems is provided by

Nekhoroshev's theorem (see, e.g., Arnol'd *et al.* (1997)), which guarantees, under smallness requirements, the stability of the motions on an open set of initial conditions for exponentially long times. Consider a Hamiltonian function of the form

$$H(y, x) = h(y) + \varepsilon f(y, x), \quad (y, x) \in B \times T^n \quad [9]$$

where B is an open subset of \mathbb{R}^n . We assume that h and f are analytic functions and that the integrable Hamiltonian h satisfies a geometric condition, called steepness. We remark that functions such as $h(L, G)$ in [8] satisfy the steepness condition. For sufficiently small values of ε , Nekhoroshev's theorem states that any motion $(y(t), x(t))$ satisfying Hamilton's equations associated with [9] is bounded for a finite (but exponentially long) time, that is,

$$\|y(t) - y(0)\| \leq y_0 \varepsilon^a, \quad \text{for } |t| \leq t_0 e^{(\varepsilon_0/\varepsilon)^b}$$

where $y_0, t_0, \varepsilon_0, a$, and b are suitable positive constants.

Nekhoroshev's theorem can be conveniently applied to the three-body problem, where it provides a confinement of the action variables, representing the semimajor axis and the eccentricity of the osculating orbit. Interesting applications of Nekhoroshev's theorem concern the investigation of the triangular Lagrangian points in the spatial, restricted three-body problem. (The Lagrangian points are five equilibrium positions of the planar, restricted three-body problem in a synodic reference frame, which rotates with the angular velocity of the primaries. Two of such positions are called triangular, since the configuration of the three bodies is an equilateral triangle in the orbital plane.) Effective estimates were developed by A Giorgilli and C Skokos, showing the existence of a stability region around the Lagrangian point L_4 , large enough to include some known asteroids. In the same framework, the exponential stability was proven by G Benettin, F Fassó, and M Guzzo for all values of the mass-ratio parameter, except for a few values of the reduced mass μ up to $\mu \simeq 0.038$.

Numerical Results

The study of the stability of the N -body problem can be investigated by performing numerical integrations of the equations of motion. The dynamics of the outer planets of the solar system (from Jupiter to Pluto) has been explored by Sussman and Wisdom (1992) using a dedicated computer, the Digital Orrery. The integration of the equations of motion was performed over 845 million years; the results provided evidence of the stability of the major planets and a chaotic behavior of Pluto. An

alternative approach, based on an average of the equations of motion over fast angles, was adopted by Laskar (1995), where the perturbing function of the spatial problem was expanded up to the second order in the masses and up to the fifth powers of the eccentricity and the inclination. The dynamics of all planets (excluding Pluto) was investigated by means of frequency analysis over a time span ranging from -15 Gyr to $+10$ Gyr. The numerical integrations provided evidence of the regularity of the external planets (from Jupiter to Neptune), a moderate chaotic behavior of Venus and the Earth, and a marked chaotic dynamics of Mercury and Mars. The computations show that the inner solar system is chaotic, with a Lyapunov time of ~ 5 Myr, thus preventing any prediction of the evolution over 100 Myr.

The Spin-Orbit Problem

The dynamics of the bodies of the solar system results from a combination of a revolutionary motion around a primary body and a rotation about an internal axis. A simple mathematical model describing the spin-orbit interaction can be introduced as follows. Let S be a triaxial ellipsoidal satellite, which moves about a central planet P . We denote by T_{rev} and T_{rot} the periods of revolution and rotation. A $p:q$ spin-orbit resonance occurs if

$$\frac{T_{\text{rev}}}{T_{\text{rot}}} = \frac{p}{q}, \quad \text{for } p, q \in \mathbb{N}, \quad q \neq 0$$

Whenever $p=q=1$, the satellite always points the same face to the host planet. Most of the evolved satellites or planets are trapped in a $1:1$ resonance, with the only exception of Mercury, which is observed in a nearly $3:2$ resonance. In order to introduce a simple mathematical model which describes the spin-orbit interaction, we assume that:

1. the satellite moves on a Keplerian orbit around the planet (with semimajor axis a and eccentricity e);
2. the spin axis is perpendicular to the orbit plane;
3. the spin axis coincides with the shortest physical axis; and
4. dissipative effects as well as perturbations due to other planets or satellites are neglected.

We denote by $A < B < C$ the principal moments of inertia of the satellite and by r and f , respectively, the instantaneous orbital radius and the true anomaly of the Keplerian orbit. Let x be the angle between the longest axis of the ellipsoid and a preassigned reference line. From standard Euler's

equations for rigid body, the equation of motion in normalized units (i.e., assuming that the period of revolution is 2π) takes the form

$$\ddot{x} + \frac{\varepsilon}{r^3} \sin(2x - 2f) = 0 \quad [10]$$

where $\varepsilon \equiv \frac{3}{2}(B - A)/C$. This equation is integrable whenever $A=B$ or in the case of zero orbital eccentricity. Due to the assumption of Keplerian motion, both r and f are known functions of the time. Therefore, we can expand [10] in Fourier series as

$$\ddot{x} + \varepsilon \sum_{m \neq 0, m=-\infty}^{\infty} W\left(\frac{m}{2}, e\right) \sin(2x - mt) = 0 \quad [11]$$

where the coefficients $W(m/2, e)$ decay as $W(m/2, e) \propto e^{|m-2|}$. A further simplification of the model is obtained as follows. According to (4), we neglected the dissipative forces and perturbations due to other bodies. The most important contribution is due to the nonrigidity of the satellite, provoking a tidal torque caused by the internal friction. The size of the dissipative effects is significantly small compared to the gravitational terms. Therefore, we decide to retain in [11] only those terms which are of the same order or larger than the average effect of the tidal torque. The following equation results:

$$\ddot{x} + \varepsilon \sum_{m \neq 0, m=N_1}^{N_2} \tilde{W}\left(\frac{m}{2}, e\right) \sin(2x - mt) = 0 \quad [12]$$

where N_1 and N_2 are suitable integers, which depend on the physical and orbital parameters of the satellite, while $\tilde{W}(m/2, e)$ are suitable truncations of the coefficients $W(m/2, e)$. We remark that eqn [12] can be derived from Hamilton's equations associated with a one-dimensional, time-dependent, nearly integrable Hamiltonian function with perturbing parameter ε and a trigonometric disturbing function.

Analytical Results

The phase space associated with [12] admits a Poincaré map showing a pendulum-like structure: the periodic orbits are surrounded by librational curves and the chaotic separatrix divides the librational regime from the region where rotational motions can take place. The three-dimensional phase space is separated by KAM rotational tori into invariant regions, providing a strong stability property for all motions confined between any pair of KAM rotational tori. Let us denote by $\mathcal{P}(p/q)$ a periodic orbit associated with the $p:q$ resonance; in the context of the model associated with [12], the

stability of the periodic orbit $\mathcal{P}(p/q)$ is obtained by showing the existence of two invariant tori $\mathcal{T}(\omega_1)$ and $\mathcal{T}(\omega_2)$ with $\omega_1 < p/q < \omega_2$. A refined computer-assisted KAM theorem has been implemented (Celletti 1990) with the aim of proving the existence of trapping invariant surfaces. Realistic estimates, in agreement with the physical values of the parameters (namely, the equatorial oblateness ε and the eccentricity e), have been obtained in several examples of spin-orbit commensurabilities, like the 1:1 Moon–Earth interaction or the 3:2 Mercury–Sun resonance.

Concerning Nekhoroshev-type estimates, the classical D'Alembert problem has been studied by Biasco and Chierchia (2002). In particular, an equatorially symmetric oblate planet moving on a Keplerian orbit around a primary body has been investigated; the model does not assume any further constraint on the spin axis. Although the Hamiltonian describing this model is properly degenerate, it is shown that Nekhoroshev-like results apply to the D'Alembert problem in the proximity of a 1:1 resonance.

Numerical Results

The model introduced in [10]–[12] often represents an unrealistic simplification of the spin-orbit dynamics. In particular, assumption (1) implies that secular perturbations of the orbital parameters are neglected, whereas the hypothesis (2) corresponds to disregarding the spin-orbit obliquity, namely the angle formed by the rotational axis with the normal to the orbital plane. Due to the presence of an equatorial bulge, the gravitational attraction of the other bodies of the solar system induces a torque, resulting in a precessional motion. It is also important to take into account the changes of the obliquity angle, whose variations might affect the climatic behavior.

A realistic model for the precession and the variation of the obliquity has been presented by Laskar (1995). The numerical simulations and the frequency-map analysis show that the Earth's obliquity is actually stable, although a large chaotic region is found in the interval between

60° and 90°. Since the present obliquity of the Earth amounts to $\sim 23.3^\circ$, the Earth is outside the dangerous region. An interesting simulation was performed to evaluate the role played by the Moon. Without the Moon, the extent of the chaotic region would greatly increase, eventually preventing the birth of an evolved life. Among the inner planets, Mars' obliquity shows larger chaotic extent, which drives to variations from 0° to 60° in a few million years. On the contrary, the external planets do not show significant chaotic regions and their obliquities are essentially stable.

See also: Averaging Methods; Dynamical Systems in Mathematical Physics: An Illustration from Water Waves; Gravitational N -Body Problem (Classical); Hamiltonian Systems: Stability and Instability Theory; KAM Theory and Celestial Mechanics; Multiscale Approaches; Stability Theory and KAM.

Further Reading

- Arnol'd VI, Kozlov VV, and Neishtadt AI (1997) *Mathematical Aspects of Classical and Celestial Mechanics*. Berlin: Springer.
- Biasco L and Chierchia L (2002) Nekhoroshev stability for the D'Alembert problem of celestial mechanics. *Atti Accademia Nazionale Lincei Classe Scienze Fisiche Matematiche Naturali Rendiconti Lincei* (9) *Matematica Applicata* 13(2): 85–89.
- Celletti A (1990) Analysis of resonances in the spin-orbit problem in celestial mechanics: the synchronous resonance (Part I). *Journal of Applied Mathematics and Physics (ZAMP)* 41: 174–204.
- Celletti A and Chierchia L (1995) *A Constructive Theory of Lagrangian Tori and Computer-Assisted Applications*. Dynamics Reported, New Series, vol. 4, pp. 60–129. Berlin: Springer.
- Delaunay C (1860) *Théorie du Mouvement de la Lune*, Mémoires de l'Académie des Sciences 1, Tome XXVIII, Paris.
- Laskar J (1995) Large scale chaos and marginal stability in the solar system. In: *XIth ICMP Meeting (Paris, July 1994)*, pp. 75–120. Cambridge, MA: International Press.
- Poincaré H (1892) *Les Méthodes Nouvelles de la Mécanique Céleste*. Paris: Gauthier-Villars.
- Sussman GJ and Wisdom J (1992) Chaotic evolution of the solar system. *Science* 241: 56–62.
- Szebehely V (1967) *Theory of Orbits*. New York: Academic Press.

Stability Theory and KAM

G Gentile, Università degli Studi "Roma Tre," Rome, Italy

© 2006 Elsevier Ltd. All rights reserved.

Introduction

A Hamiltonian system is a dynamical system whose equations of motions can be written in terms of a scalar function, called the Hamiltonian of the system: if one uses coordinates (p, q) in a domain (phase space) $\mathcal{D} \subset \mathbb{R}^{2N}$, where N is the number of independent variables one needs to identify a configuration of the system (degrees of freedom), there is a function $\mathcal{H}(p, q)$ such that $\dot{p} = -\partial\mathcal{H}/\partial q$ and $\dot{q} = \partial\mathcal{H}/\partial p$. An integrable (Hamiltonian) system is a Hamiltonian system which, in suitable coordinates $(A, \alpha) \in \mathcal{A} \times T^N$, where \mathcal{A} is an open subset of \mathbb{R}^N and $T = \mathbb{R}/2\pi\mathbb{Z}$ is the standard torus, can be described by a Hamiltonian $\mathcal{H}_0(A)$, that is, depending only on A . The coordinates (A, α) are called action-angle variables. In such a case the dynamics is trivial: any initial condition (A_0, α_0) evolves in such a way that the action variables are constants of motion (i.e., $A(t) = A_0$ for all $t \in \mathbb{R}$), while the angles grow linearly in time as $\alpha(t) = \alpha_0 + \omega t$, where $\omega = \omega(A_0) \equiv \partial_A \mathcal{H}_0(A_0)$ is called the rotation (or frequency) vector. An integrable system can be thought of as a collection of decoupled (i.e., independent) rotators: the entire phase space $\mathcal{A} \times T^N$ is foliated into invariant tori and all motions are quasiperiodic. Integrable systems are stable, in the sense that nearby initial conditions separate at most linearly in time (in particular, the actions do not separate at all): mathematically, this is expressed by the fact that all the Lyapunov exponents are nonpositive.

An example of an integrable system is any one-dimensional conservative mechanical system, in any region of phase space in which motions are bounded. By increasing the number of degrees of freedom, exhibiting nontrivial integrable systems can become a difficult task. The problem of studying the effects of even small Hamiltonian perturbations on integrable systems and of understanding if the latter remain stable, in the aforementioned sense, was considered by Poincaré to be the fundamental problem of dynamics. For a long time, it was commonly thought that all motions could be reduced to superpositions of periodic motions, hence to quasiperiodic motions, but at the end of nineteenth century it was realized by Boltzmann and Poincaré that such a picture was too naive, and that in reality more complicated motions were possible.

As a consequence of this, it became a widespread belief that, even when starting from an integrable system, the introduction of an arbitrarily small perturbation would break integrability.

This belief was strengthened by the work of Poincaré (1898), who showed that the series describing the solution in a perturbation theory approach are in general divergent. The source of divergence in perturbation series is the presence of small divisors, that is, of denominators of the kind of $\omega \cdot \nu$, where ω is the rotation vector that should characterize the invariant torus (if existent) and ν is any integer vector. Despite this, however, perturbation series (known as Lindstedt series) continued to be extensively used by astronomers in problems of celestial mechanics, such as the study of planetary motions, for the simple reason that they provided predictions in good agreement with the observations. But the feeling that the underlying mathematical tools were unsatisfactory persisted.

In fact, the well-known Fermi–Pasta–Ulam numerical experiment, in 1955, was originally conceived in the spirit of confirming that integrability would in general be easily lost. Consider a chain with N harmonic oscillators, with, say, periodic boundary conditions, coupled with cubic and quartic two-body potentials, so that the Hamiltonian is

$$\begin{aligned} \mathcal{H}(p, q) &= \sum_{i=1}^N \frac{1}{2} p_i^2 + W(q_{i+1} - q_i) \\ W(x) &= \frac{1}{2} x^2 + \frac{\alpha}{3} x^3 + \frac{\beta}{4} x^4 \end{aligned} \quad [1]$$

for α, β real parameters and $(p, q) \in \mathbb{R}^N \times \mathbb{R}^N$. One can introduce new variables such that the Hamiltonian, for $\alpha = \beta = 0$, can be written as

$$\mathcal{H}_0(A) = \sum_{i=1}^N \frac{1}{2} (P_k^2 + \omega_k Q_k^2) = \omega \cdot A \quad [2]$$

for a suitable rotation vector $\omega = (\omega_1, \dots, \omega_N) \in \mathbb{R}^N$ (an explicit computation gives $\omega_k = 2 \sin(k\pi/N)$).

Consider an initial condition in which all the energy is confined to a few modes, that is, $A_k \neq 0$ at $t=0$ only for a few values of k . For $\alpha = \beta = 0$, the system is integrable, so that $A_k(t) = 0$ for all $t \in \mathbb{R}$ and for all k such that $A_k(0) = 0$. If the system ceases to be integrable when the perturbation is switched on, the energy is likely to start to be shared among the various modes, and after a long enough time has

elapsed, an equidistribution of the energy among all modes (thermalization) might be expected. At least this behavior was expected by Fermi, Pasta, and Ulam, but it was not what they found numerically: on the contrary, all the energy seemed to remain associated with the modes close to the few initially excited ones.

At about the same time, Kolmogorov (1954) published a breakthrough paper going exactly in the opposite direction: if one perturbs an integrable system, under some mild conditions on the integrable part, most of the tori are preserved, although slightly deformed. A more precise statement is the following.

Theorem 1 *Let an N -degree-of-freedom Hamiltonian system be described by an analytic Hamiltonian of the form*

$$\mathcal{H}(\mathbf{A}, \boldsymbol{\alpha}) = \mathcal{H}_0(\mathbf{A}) + \varepsilon f(\mathbf{A}, \boldsymbol{\alpha}) \quad [3]$$

with ε a real parameter (perturbation parameter), f a 2π -periodic function of each angle variable (potential or perturbation), and $\mathcal{H}_0(\mathbf{A})$ satisfying the nondegeneracy condition $\det \partial_{\mathbf{A}}^2 \mathcal{H}_0(\mathbf{A}) \neq 0$ (anisochrony condition). If $\boldsymbol{\omega} = \boldsymbol{\omega}(\mathbf{A}) \equiv \partial_{\mathbf{A}} \mathcal{H}_0(\mathbf{A})$ is fixed to satisfy the Diophantine condition

$$|\boldsymbol{\omega} \cdot \mathbf{v}| > \frac{C_0}{|\mathbf{v}|^\tau} \quad \forall \mathbf{v} \in \mathbb{Z}^N \setminus \mathbf{0} \quad [4]$$

for some constants $C_0 > 0$ and $\tau > N - 1$ (here $|\mathbf{v}| = |\nu_1| + \dots + |\nu_N|$ and \cdot denotes the standard inner product: $\boldsymbol{\omega} \cdot \mathbf{v} = \omega_1 \nu_1 + \dots + \omega_N \nu_N$), then there is an invariant torus with rotation vector $\boldsymbol{\omega}$ for ε small enough, say for ε smaller than some value ε_0 depending on C_0 and τ (and on the function f).

By saying that there is an invariant torus with rotation vector $\boldsymbol{\omega}$, one means that there is an invariant surface in phase space on which, in suitable coordinates, the dynamics is the same as in the unperturbed case, and the conjugation (i.e., the change of variables which leads to such coordinates) is analytic in the angle variables and in the perturbation parameter. One also says that the torus of an integrable system ($\varepsilon = 0$) is preserved (or even persists) under a small perturbation.

Note that, *a posteriori*, this proves convergence of the perturbation series: however, a direct check of convergence was performed only recently by Eliasson (1996). Kolmogorov's proof was based on a completely different idea, that is, by performing iteratively a sequence of canonical transformations (which are changes of coordinates preserving the

Hamiltonian structure of the equations of motion) such that at each step the size of the perturbation is reduced. Of course, on the basis of Poincaré's result, this iterative procedure cannot work for all initial conditions (e.g., when $\boldsymbol{\omega}$ does not satisfy [4]). The key point in Kolmogorov's scheme is to fix the rotation vector $\boldsymbol{\omega}$ of the torus one is looking for, in such a way that the small divisors are controlled through the Diophantine condition [4] and the exponentially fast convergence of the algorithm.

New proofs and extensions of Kolmogorov's theorem were given later by Arnol'd (1962) and by Moser (1962); hence, the acronym KAM to denote such a theorem. Arnol'd gave a more detailed (and slightly different) proof compared to the original one by Kolmogorov, and applied the result to the planar three-body problem, thus showing that physical applications of the theorem were possible. Moser, on the other hand, proposed a modified method using a technique introduced by Nash (which approximates smooth functions with analytical ones) to deal with the case of systems with finite smoothness.

For fixed small enough ε , the surviving invariant tori cover a large portion of the phase space, called the Kolmogorov set; the relative measure of the region of phase space which is not filled by such tori tends to zero at least as $\sqrt{\varepsilon}$ for $\varepsilon \rightarrow 0$. A system described by a Hamiltonian like [3] is then called a quasi-integrable Hamiltonian system.

The excluded region of phase space corresponds to the unperturbed tori which are destroyed by the perturbation: the rotation vectors of such tori are close to a resonance, that is, to a value $\boldsymbol{\omega}$ such that $\boldsymbol{\omega} \cdot \mathbf{v} = 0$ for some integer vector \mathbf{v} , and these are exactly the vectors which do not satisfy the Diophantine condition [4] for any value C_0 . A subset of phase space of this kind is called a resonance region.

At first sight, this would seem to provide an explanation for the results found by Fermi, Pasta, and Ulam, but this is not quite the case. First, the threshold value ε_0 depends on N , and goes to zero very fast as $N \rightarrow \infty$ (in general as $N!^{-\alpha}$ for some $\alpha > 0$); however, the results of the numerical experiments apparently were insensitive to the number N of oscillators. Second, the KAM theorem deals with maximal tori, that is, tori characterized by rotation vectors which have as many components as the number of degrees of freedom, while the rotation vectors of the numerical quasiperiodic solutions seem to involve just a small number of components.

Finally, as an extra problem, the validity of the nondegeneracy condition for the unperturbed

Hamiltonian is violated, because the unperturbed Hamiltonian is linear in the action variables (one says that the Hamiltonian is isochronous). Recently, Rink (2001), by continuing the work by Nishida, showed that in the Fermi–Pasta–Ulam problem it is possible to perform a canonical change of coordinates such that in the new variables the Hamiltonian becomes anisochronous: one uses part of the perturbation to remove isochrony. But the other two obstacles remain.

Lower-Dimensional Tori

A natural question is what happens to the invariant tori corresponding to rotation vectors which are not rationally independent, that is, vectors satisfying n resonance conditions, such as $\omega \cdot \mathbf{v}_i = 0$ for n independent vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$, with $1 \leq n \leq N - 2$ (the case $n = N - 1$ corresponds to periodic orbits and is comparatively easy); for instance, one can take $\omega = (\omega_1, \dots, \omega_n, 0, \dots, 0)$ and, by a suitable linear change of coordinates, one can always make the reduction to a case of this kind. In particular, one can ask if a result analogous to the KAM theorem holds for these tori. Such a problem for the model [3] has not been studied very widely in the literature. What has usually been considered is a system of n rotators coupled with a system with $s = N - n$ degrees of freedom near an equilibrium point: then one calls normal coordinates the coordinates describing the latter, and the role of the parameter ε is played by the size of the normal coordinates (if their initial conditions are chosen near the equilibrium point). In the absence of perturbation (i.e., for $\varepsilon = 0$), one has either hyperbolic or elliptic or, more generally, mixed tori, according to the nature of the equilibrium points: one refers to these tori as lower-dimensional tori, as they represent n -dimensional invariant surfaces in a system with N degrees of freedom. Then one can study the preservation of such tori.

One can prove that, in such a case, at least if certain generic conditions are satisfied, in suitable coordinates, n angles rotate with frequencies $\omega_1, \dots, \omega_n$, respectively, while the remaining $N - n$ angles have to be fixed close to some values corresponding to the extremal points of the function obtained by averaging the potential over the rotating angles.

The case of hyperbolic tori is easier, as in the case of elliptic tori one has to exclude some values of ε to avoid some further resonance conditions between the rotation vector ω and the normal frequencies λ_k (i.e., the eigenvalues of the linearized system

corresponding to the normal coordinates), known as the first and second Mel'nikov conditions:

$$\begin{aligned} |\omega \cdot \mathbf{v} \pm \lambda_k| &> \frac{C_0}{|\mathbf{v}|^\tau} \quad \forall \mathbf{v} \in \mathbb{Z}^N \setminus \mathbf{0}, \quad \forall 1 \leq k \leq s \\ |\omega \cdot \mathbf{v} \pm \lambda_k \pm \lambda_{k'}| &> \frac{C_0}{|\mathbf{v}|^\tau} \quad \forall \mathbf{v} \in \mathbb{Z}^N \setminus \mathbf{0} \\ \forall 1 \leq k, k' \leq s \end{aligned} \quad [5]$$

Such conditions appear, with the values of the normal frequencies slightly modified by terms depending on ε , at each iterative step, and at the end only for values of ε belonging to some Cantor set one can have elliptic lower-dimensional tori.

The second Mel'nikov conditions are not really necessary, and in fact they can be relaxed as Bourgain (1994) has shown; this is an important fact, as it allows degenerate normal frequencies, which were forbidden in the previous works by Kuksin (1987), Eliasson (1988), and Pöschel (1989).

Similar results also apply in the case of lower-dimensional tori for the model [3], which represents sort of a degenerate situation, as the normal frequencies vanish for $\varepsilon = 0$. Again, one has to use part of the perturbation to remove the complete degeneracy of normal frequencies.

Quasiperiodic Solutions in Partial Differential Equations

For explaining the Fermi–Pasta–Ulam experiment, one has to deal with systems with arbitrarily many degrees of freedom. Hence, it is natural to investigate systems which have *ab initio* infinitely many degrees of freedom, such as the nonlinear wave equation, $u_{tt} - u_{xx} + V(x)u = \varphi(u)$, the nonlinear Schrödinger equation, $iu_t - u_{xx} + V(x)u = \varphi(u)$, the nonlinear Korteweg–de Vries equation $u_t + u_{xxx} - 6u_x u = \varphi(u)$, and other systems of nonlinear partial differential equations (PDEs); the continuum limit of the Fermi–Pasta–Ulam model gives indeed a nonlinear Korteweg–de Vries equation, as shown by Zabuski and Kruskal (1965). Here $(t, x) \in \mathbb{R} \times [0, \pi]^d$, if d is the space dimension, and either periodic ($u(0, t) = u(\pi, t)$) or Dirichlet ($u(0, t) = u(\pi, t) = 0$) boundary conditions can be considered; $\varphi(u)$ is a function analytic in u and starting from orders strictly higher than one, while $V(x)$ is an analytic function of x , depending on extra parameters ξ_1, \dots, ξ_n . Such a function is introduced essentially for technical reasons, as we shall see that the eigenvalues λ_k of the Sturm–Liouville operator $-\partial_x^2 + V(x)$ must satisfy some Diophantine conditions. If we set $V(x) = \mu \in \mathbb{R}$ in the nonlinear wave equation, we obtain the Klein–Gordon equation, which, in the particular case $\mu = 0$,

reduces to the string equation. Again, the role of the perturbation parameter is played by the size of the solution itself.

Small-amplitude periodic and quasiperiodic solutions for PDE systems have been extensively studied, among others, by Kuksin, Wayne, Craig, Pöschel, and Bourgain. Results for such systems read as follows. Consider for concreteness the one-dimensional nonlinear wave equation with Dirichlet boundary conditions and with $\varphi(u) = u^3 + O(u^5)$. When the nonlinear function $\varphi(u)$ is absent, any solution of the linear wave equation $u_{tt} - u_{xx} + V(x)u = 0$ is a superposition of either finitely or infinitely many periodic solutions with frequencies λ_k determined by the function $V(x)$. Let $u_0(\omega t, x)$ be a quasiperiodic solution of the linear wave equation with rotation vector $\omega \in \mathbb{R}^n$, where $\omega_k = \lambda_{m_k}$, for some n -tuple $\{m_1, \dots, m_n\}$. Then for ε small enough there exists a subset Ξ_ε of the space of parameters with large Lebesgue measure (more precisely, with complementary Lebesgue measure which tends to zero when $\varepsilon \rightarrow 0$) such that for all $\xi = (\xi_1, \dots, \xi_n) \in \Xi_\varepsilon$ there is a solution $u_\varepsilon(t, x)$ of the nonlinear wave equation and a rotation vector ω_ε satisfying the conditions

$$\begin{aligned} |u_\varepsilon(t, x) - \sqrt{\varepsilon} u_0(\omega_\varepsilon t, x)| &\leq C\varepsilon \\ |\omega_\varepsilon - \omega| &< C\varepsilon \end{aligned} \quad [6]$$

for some positive constant C .

The case $n = 1$ (periodic solutions) is not as easy as the finite-dimensional case, because there are infinitely many normal frequencies, so that there are small divisor problems which for finite-dimensional systems appear only for $n \geq 2$.

For the nonlinear wave equation and the Schrödinger equation, if $n \geq 1$, one can take $V(x) = \mu$, but one needs $\mu \neq 0$; for $n > 1$, one can take $V(x) = \mu$, as one can perform a preliminary transformation leading to an equation in which a function depending on parameters naturally appears, as shown by Kuksin and Pöschel (1996). For $n = 1$, the case $\mu = 0$ has been very recently solved by Gentile *et al.* (2005).

Statements for more general situations can also be obtained, while extensions to space dimensions $d \geq 2$ are not trivial and have been obtained only recently by Bourgain (1998). The above result also holds if the number of components of the rotation vector is less than the number of parameters: one uses such parameters because one needs to impose some Diophantine conditions such as [5], now for all the frequencies $\lambda_k = \omega_k, k \notin \{m_1, \dots, m_n\}$. Again, the second Mel'nikov conditions were shown by Bourgain to be unnecessary, and this is an essential ingredient for the higher-dimensional case.

Even if systems of the type considered above have been widely studied, they remain significantly different from a discrete system such as the chain of oscillators [1] for N large enough (also in the limit $N \rightarrow \infty$), so that the results which have been found for PDE systems do not really provide an explanation for the numerical findings.

Also in the case of lower-dimensional tori for finite-dimensional systems the main problem is that, even if such tori exist, it is not clear what relevance they can have for the dynamics (a case in which hyperbolic tori play a role is considered later). An important feature of maximal tori is that they fill most of the phase space, a property which certainly does not hold for lower-dimensional tori, which lie outside the Kolmogorov set.

In the Fermi–Pasta–Ulam experiment, one considers initial conditions close to lower-dimensional tori; hence, an interesting problem is to study their stability, that is, how fast the trajectories starting from such initial conditions drift away.

Arnol'd Diffusion and Nekhoroshev's Theorem

Consider again the maximal tori. For $N = 2$, the preservation of most of the invariant tori prevents the possibility of diffusion in phase space: the tori represent two-dimensional surfaces in a three-dimensional space (as dynamics occur on the level surfaces of the energy in a four-dimensional space), so that, if an initial condition is trapped in a gap between two tori, the corresponding trajectory remains confined forever between them. The situation is quite different for $N \geq 3$: in such a case, the tori do not represent a topological obstruction to diffusion any more.

That mechanisms of diffusion are really possible was shown by Arnol'd (1963). Because of the perturbation, lower-dimensional hyperbolic tori appear inside the resonance regions, with their stable and unstable manifolds (whiskers). It is possible that these manifolds of the same torus intersect with a nonvanishing angle (homoclinic angle); as a consequence, the angles between the stable and unstable manifolds of nearby tori (heteroclinic angles) can also be different from zero, and one can find a set of hyperbolic lower-dimensional tori such that the unstable manifold of each of them intersects the stable manifold of the torus next to it: one says that such tori form a transition chain of heteroclinic connections. Then there can be trajectories moving along such connections, producing at the end a drift of order 1 (in ε) in the action variables. Such a phenomenon is referred to as Arnol'd diffusion.

Of course, diffusing trajectories should be located in the region of phase space where there are no invariant tori (hence, a very small region when ε is small), but an important consequence is that, unlike what happens in the unperturbed case, not all motions are stable: in particular, the action variables can change by a large amount over long times.

Providing interesting examples of Hamiltonian systems in which Arnol'd diffusion can occur is not so easy: in fact, for the diffusion to really occur, one needs a lower bound on the homoclinic angles, and to evaluate these angles can be difficult. For instance, Arnold's (1963) original example, which describes a system near a resonance region, is a two-parameter system given by

$$\begin{aligned} \frac{1}{2}(A_1^2 + A_2^2) + A_3 + \mu(\cos \alpha_1 - 1) \\ + \varepsilon\mu(\cos \alpha_1 - 1)(\sin \alpha_2 + \cos \alpha_3) \end{aligned} \quad [7]$$

and the angles can be proved to be bounded from below only by assuming that the perturbation parameter ε is exponentially small with respect to the other parameter μ , which in turn implies a situation not really convincing from a physical point of view. More generally, for all the examples which are discussed in literature, the relation with physics (as the d'Alembert problem on the possibility for a planet to change the inclination of the precession cone) is not obvious.

So the question naturally arises as to how fast can such a mechanism of diffusion be, and how relevant is it for practical purposes. A first answer is provided by a theorem of Nekhoroshev (1977), which states the following result.

Theorem 2 *Suppose we have an N -degree-of-freedom quasi-integrable Hamiltonian system, where the unperturbed Hamiltonian satisfies some condition such as convexity (or a weaker one, known as steepness, which is rather involved, to state in a concise way); for concreteness consider a function $\mathcal{H}_0(\mathbf{A})$ in [2] which is quadratic in \mathbf{A} . Then there are two positive constants a and b such that for times t up to $O(\exp(\varepsilon^{-b}))$ the variations of the action variables cannot be larger than $O(\varepsilon^a)$.*

The constants a and b depend on N , and they tend to zero when $N \rightarrow \infty$; Lochak and Neishtadt (1992) and Pöschel (1993) found estimates $a = b = 1/2N$, which are probably in general optimal. Nekhoroshev's theorem is usually stated in the form above, but it provides more information than that explicitly written: the trajectories, when trapped into a resonance region, drift away and come close to some invariant torus, and then they behave like quasiperiodic motions, up to very small corrections, for a long time, until they enter some other

resonance region, and so on. Of course, for initial conditions on some invariant torus, KAM theorem applies, but the new result concerns initial conditions which do not belong to any tori.

Nekhoroshev's theorem gives a lower bound for the diffusion time, that is, the time required for a drift of order 1 to occur in the action variables. But, of course, an upper bound would also be desirable. The diffusion times are related to the amplitude of the homoclinic angles, which are very small (and difficult to estimate as stated before). The strongest results in this direction have been obtained with variational methods, for instance, by Bessi, Bernard, Berti, and Bolle: at best, for the diffusion time, one finds an estimate $O(\mu^{-1} \log \mu^{-1})$, if μ is the amplitude of the homoclinic angles (which in turn are exponentially small in some power of ε , as one can expect as a consequence of Nekhoroshev's theorem).

Then one can imagine that the results of the Fermi-Pasta-Ulam experiment can also be interpreted in the light of Nekhoroshev's theorem. The solutions one finds numerically certainly do not correspond to maximal tori, but one could expect that they could be solutions which appear to be quasiperiodic for long but finite times (e.g., moving near some lower-dimensional torus determined by the initial conditions), and that if one really insists on observing the time evolution for a very long time, then deviations from quasiperiodic behavior could be detected. This is an appealing interpretation, and the most recent numerical results make it plausible: Galgani and Giorgilli (2003) have found numerically that the energy, even if initially confined to the lower modes, tend to be shared among all the other modes, and higher the modes the longer is the time needed for the energy to flow to them. Of course, this does not settle the problem, as there is still the issue of the large number of degrees of freedom; furthermore, for large N the spacing between the frequencies is small, and they become almost degenerate. Hence, the problem still has to be considered as open.

Stability versus Chaos

The main problem in applying the KAM theorem seems to be related to the small value of the threshold ε_0 which is required. In general, when the size of the perturbation parameter is very large, the region of phase space filled with invariant tori decreases (or even disappears), and chaotic motions appear. By the latter, one generally means motions which are highly sensitive to the initial conditions: a small variation of the initial conditions produces a catastrophic variation in the corresponding trajectories (this is due to the appearance of strictly positive Lyapunov exponents).

A natural question is then how such a result as the KAM theorem is meaningful in physical situations: in other words, for which systems the KAM theorem can really apply.

One of the main motivations to study such a problem was to explain astronomical observations and to study the stability of the solar system. In order to apply the KAM theorem to the solar system, one has to interpret the gravitational forces between the planets as perturbations of a collection of several decoupled two-body systems (each planet with the Sun). One can write the masses of the planets as εm_i , and ε plays the role of the perturbation parameter. The corresponding Hamiltonian (after suitable reductions and scalings) is

$$\sum_{i=1}^N \frac{p_i^2}{2\mu_i} - \sum_{i=1}^N \frac{m_i m_0}{|q_i|} + \varepsilon \sum_{1 \leq i < j \leq N} \frac{p_i \cdot p_j}{m_0} + \varepsilon \sum_{1 \leq i < j \leq N} \frac{m_i m_j}{|q_i - q_j|} \quad [8]$$

where $i=0$ corresponds to the Sun, while $i=1, \dots, N$ correspond to the planets (hence $N=9$), m_0 is the mass of the Sun, and $\varepsilon \mu_i$ are the reduced masses ($\mu_i^{-1} = m_i^{-1} + \varepsilon m_0^{-1}$); here $(q_i, p_i) \in \mathbb{R}^3 \times \mathbb{R}^3$, $i=0, \dots, N$, the inner product in $p_i \cdot p_j$ is in \mathbb{R}^3 , and the norm $|\cdot|$ is the Euclidean one.

A first difficulty is that the solar system is a properly degenerate system; that is, the unperturbed Hamiltonian does not depend on all the action variables. But such a degeneracy can be removed by performing a canonical change of coordinates which produces a new Hamiltonian in which the integrable part contains new terms of order ε depending on all action variables and is nondegenerate, while the perturbation becomes of order ε^2 : the angle variables corresponding to the actions not originally appearing in the unperturbed Hamiltonian are called the slow variables, while the others are called the fast variables.

However, a naive implementation of the KAM theorem, in general, even for simplified but still realistic systems, would provide a preposterously small value of the threshold ε_0 . The problem could be just a computational one: in principle, a very refined estimate of the threshold could give a better value, so that it is very difficult to decide analytically if the real values of the planetary masses allow the solar system to fall inside the regime of applicability of the KAM theorem. Results in this direction have been obtained, but only for special situations: for instance, by considering the restricted planar circular three-body problem (which provides a simplified description of the system "Sun + Jupiter + asteroid"), Celletti and Chierchia

(1997) found analytical bounds on the perturbation parameters comparable with the physical values. Of course, this is not at all conclusive for the general situation in which all planets (with their satellites and the asteroids) are considered together; in particular, it does not shed light on the problem of the stability of the entire solar system.

On the contrary, extensive numerical simulations performed by Laskar (starting from 1989) seem to suggest that the solar system is unstable. Deflections from the current orbits could be produced to such an extent that collisions between planets could not be avoided: Mercury could collide with Venus and be ejected from the solar system. An important issue is to consider the times over which such phenomena can occur. Laskar's numerical simulations show that such times are less than the estimated age of the solar system, and that one can make accurate predictions for the planetary motions only for a finite amount of time (~ 100 Myr). Furthermore, the assumed partial instability of the solar system has also been used by Laskar (2004) to explain some observed phenomena such as the evolution of the obliquity (which is the angle between equator and orbital plane) of some planets. Of course, these simulations have been carried out with several approximations, as that of averaging over the fast variables, which allows one to use a large integration step in the numerical integration of the equations of motion for the resulting system. This is the so-called secular system introduced by Lagrange: instead of the fast motion of the planets, one describes the slow deformations of the planetary orbits (imagining the planets as regions of mass spread along their orbits).

See also: Averaging Methods; Bifurcation Theory; Billiards in Bounded Convex Domains; Diagrammatic Techniques in Perturbation Theory; Dynamical Systems and Thermodynamics; Gravitational N -Body Problem (Classical); Hamiltonian Systems: Stability and Instability Theory; Hamilton–Jacobi Equations and Dynamical Systems: Variational Aspects; Integrable Systems and Discrete Geometry; KAM Theory and Celestial Mechanics; Localization for Quasiperiodic Potentials; Stability Problems in Celestial Mechanics; Synchronization of Chaos; Weakly Coupled Oscillators.

Further Reading

- Arnol'd VI (1963) Proof of a theorem of A. N. Kolmogorov on the preservation of conditionally periodic motions under a small perturbation of the Hamiltonian. *Russian Mathematical Surveys* 18: 85–192.
- Arnol'd VI (1964) Instability of dynamical systems with many degrees of freedom. *Soviet Mathematics Doklady* 5: 581–585.

- Arnol'd VI, Kozlov VV, and Neishtadt AI (1988) *Dynamical Systems III*. Encyclopedia of Mathematical Sciences, vol. 3. Berlin: Springer.
- Bourgain J (1994) Construction of quasi-periodic solutions for Hamiltonian perturbations of linear equations and applications to nonlinear PDE. *International Mathematics Research Notices* 1994(11), 475–497.
- Bourgain J (1998) Quasi-periodic solutions of Hamiltonian perturbations of 2D linear Schrödinger equations. *Annals of Mathematics* 148: 363–439.
- Bourgain J (2005) *Green's Function Estimates for Lattice Schrödinger Operators and Applications*. Princeton: Princeton University Press.
- Celletti A and Chierchia L (1997) On the stability of realistic three-body problems. *Communications in Mathematical Physics* 186: 413–449.
- Eliasson LH (1996) Absolutely convergent series expansions for quasi periodic motions. *Mathematical Physics Electronic Journal* 2, paper 4 (electronic). Preprint 1988.
- Eliasson LH (1988) Perturbations of stable invariant tori for Hamiltonian systems. *Annali della Scuola Normale Superiore di Pisa* 15: 115–147.
- Ford J (1992) The Fermi–Pasta–Ulam problem: paradox turns discovery. *Physics Reports* 213: 271–310.
- Galgani L and Giorgilli A (2003) Recent results on the Fermi–Pasta–Ulam problem. *Rossiiskaya Akademiya Nauk. Sankt-Peterburgskoe Otdelenie. Matematicheskii Institut im. V. A. Steklova. Zapiski Nauchnykh Seminarov (POMI)* 300: 145–154.
- Gallavotti G (1986) Quasi-integrable mechanical systems. In: *Phénomènes critiques, systèmes aléatoires, théories de jauge* (Les Houches, 1984), pp. 539–624. Amsterdam: North-Holland.
- Gallavotti G, Bonetto F, and Gentile G (2004) *Aspects of Ergodic, Qualitative and Statistical Theory of Motion*. Berlin: Springer.
- Gentile G, Mastropietro V, and Procesi M (2005) Periodic solutions for completely resonant nonlinear wave equations with Dirichlet boundary conditions. *Communications in Mathematical Physics* 257: 319–362.
- Kolmogorov AN (1954) On conservation of conditionally periodic motions for a small change in Hamilton's function. *Doklady Akademii Nauk SSSR* 98: 527–530 (Russian).
- Kuksin SB (1987) Hamiltonian perturbations of infinite-dimensional linear systems with imaginary spectrum. *Functional Analysis and its Applications* 21: 192–205.
- Kuksin SB (1993) *Nearly Integrable Infinite-Dimensional Hamiltonian Systems*, Lecture Notes in Mathematics, vol. 1556. Berlin: Springer.
- Kuksin SB and Pöschel J (1996) Invariant Cantor manifolds of quasi-periodic oscillations for a nonlinear Schrödinger equation. *Annals of Mathematics* 143: 149–179.
- Laskar J (2004) Chaos in the solar system. In: Iagolnitzer D, Rivasseau V, and Zinn-Justin J (eds.) *Proceedings of the International Conference on Theoretical Physics TH2002*, (Paris, 2002). Basel: Birkhäuser.
- Lochak P and Neishtadt AI (1992) Estimates of stability time for nearly integrable systems with a quasiconvex Hamiltonian. *Chaos* 2: 495–499.
- Moser J (1962) On invariant curves of area-preserving mappings of an annulus. *Nachrichten der Akademie der Wissenschaften in Göttingen* 1962: 1–20.
- Moser J (1973) *Stable and Random Motions in Dynamical Systems*, Annals of Mathematical Studies. Princeton: Princeton University Press.
- Nekhoroshev NN (1977) An exponential estimate of the time of stability of nearly integrable Hamiltonian systems. *Russian Mathematical Surveys* 32: 1–65.
- Pöschel J (1989) On elliptic lower-dimensional tori in Hamiltonian systems. *Mathematische Zeitschrift* 202: 559–608.
- Pöschel J (1993) Nekhoroshev estimates for quasi-convex Hamiltonian systems. *Mathematische Zeitschrift* 213: 187–216.
- Rink B (2001) Symmetry and resonance in periodic FPU chains. *Communications in Mathematical Physics* 218: 665–685.
- Zabusky NJ and Kruskal MD (1965) Interaction of “solitons” in a collisionless plasma and the recurrence of initial states. *Physical Review Letters* 15: 240–243.

Standard Model of Particle Physics

G Altarelli, CERN, Geneva, Switzerland

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The standard model (SM) is a consistent, finite, and – within the limitations of our present technical ability – computable theory of fundamental microscopic interactions that successfully explains most of the known phenomena in elementary particle physics. The SM describes strong, electromagnetic, and weak interactions. All microscopic phenomena observed to date can be attributed to one or the other of these interactions. For example, the forces that hold together the protons and the neutrons in

the atomic nuclei are due to strong interactions; the binding of electrons to nuclei in atoms or of atoms in molecules is caused by electromagnetism; and the energy production in the Sun and the other stars occurs through nuclear reactions induced by weak interactions. In principle, gravitational forces should also be included in the list of fundamental interactions but their impact on fundamental particle processes at accessible energies is totally negligible.

The structure of the SM is a generalization of that of quantum electrodynamics (QED), in the sense that it is a renormalizable field theory based on a local symmetry (i.e., separately valid at each spacetime point x) that extends the gauge invariance of electrodynamics to a larger set of

conserved currents and charges. There are eight strong charges, called “color” charges and four electroweak charges (which, in particular, include the electric charge). The commutators of these charges form the $SU(3) \otimes SU(2) \otimes U(1)$ algebra. In QED, the interaction between two matter particles with electric charges (e.g., two electrons) is mediated by the exchange of one (or more) photons emitted by one electron and reabsorbed by the second. In the SM the matter fields, all of spin $1/2$, are the quarks, the constituents of protons, neutrons, and all hadrons, endowed with both color and electroweak charges, and the leptons (the electron e^- , the muon μ^- , the tauon τ^- , plus the three associated neutrinos ν_e , ν_μ , and ν_τ) with no color but with electroweak charges. The matter fermions come in three generations or families with identical quantum numbers but different masses. The pattern is as follows:

$$\begin{bmatrix} u & u & u & \nu_e \\ d & d & d & e \\ t & t & t & \nu_\tau \\ b & b & b & \tau \end{bmatrix}, \quad \begin{bmatrix} c & c & c & \nu_\mu \\ s & s & s & \mu \end{bmatrix} \quad [1]$$

Each family contains a weakly charged doublet of quarks, in three color replicas, and a colorless weakly charged doublet with a neutrino and a charged lepton. At present, there is no explanation for this triple repetition of fermion families. The force carriers, of spin 1, are the photon γ , the weak interaction gauge bosons W^+ , W^- , and Z_0 and the eight gluons g that mediate the strong interactions. The photon and the gluons have zero masses as a consequence of the exact conservation of the corresponding symmetry generators, the electric charge and the eight color charges. The weak bosons W^+ , W^- , and Z_0 have large masses ($m_W \sim 80.4$ GeV, $m_Z = 91.2$ GeV), signaling that the corresponding symmetries are badly broken. In the SM, the spontaneous breaking of the electroweak gauge symmetry is induced by the Higgs mechanism, which predicts the presence of one (or more) spin 0 particles in the physical spectrum, the Higgs boson(s), not yet experimentally observed. A tremendous experimental effort is underway or planned to reveal the Higgs sector as the last crucial missing link in the SM verification.

Quantum Chromodynamics

The statement that quantum chromodynamics (QCD) is a renormalizable gauge theory based on

the group $SU(3)$ with color triplet quark matter fields fixes the QCD Lagrangian density to be

$$\mathcal{L} = -\frac{1}{4} \sum_{A=1}^8 F^{A\mu\nu} F_{\mu\nu}^A + \sum_{j=1}^{n_f} \bar{q}_j (i\not{D} - m_j) q_j \quad [2]$$

Here q_j are the quark fields (of n_f different flavors) with mass m_j ; $\not{D} = D_\mu \gamma^\mu$, where γ^μ are the Dirac matrices and D_μ is the covariant derivative

$$D_\mu = \partial_\mu - ie_s \sum_A t^A g_\mu^A \quad [3]$$

e_s is the gauge coupling (in analogy with QED,

$$\alpha_s = \frac{e_s^2}{4\pi} \quad [4]$$

here and throughout this article natural units, $\hbar = c = 1$, are used); g_μ^A , $A = 1, \dots, 8$, are the gluon fields, and t^A are the $SU(3)$ group generators in the triplet representation of quarks (i.e., t_A are 3×3 matrices acting on q); the generators obey the commutation relations $[t^A, t^B] = iC_{ABC}t^C$, where C_{ABC} are the complete antisymmetric structure constants of $SU(3)$ (the normalization of C_{ABC} and of e_s is specified by $\text{tr}[t^A t^B] = 1/2\delta^{AB}$);

$$F_{\mu\nu}^A = \partial_\mu g_\nu^A - \partial_\nu g_\mu^A - e_s C_{ABC} g_\mu^B g_\nu^C \quad [5]$$

The physical vertices in QCD include the gluon–quark–antiquark vertex, analogous to the QED photon–fermion–antifermion coupling, but also the three-gluon and four-gluon vertices, of order e_s and e_s^2 , respectively, which have no analog in an abelian theory like QED. In QED, the photon (a neutral particle) is coupled to all electrically charged particles. In QCD, the gluons are colored, hence self-coupled. This is reflected in the fact that in QED $F_{\mu\nu}$ is linear in the gauge field, so that the term $F_{\mu\nu}^2$ in the Lagrangian is a pure kinetic term, while in QCD $F_{\mu\nu}^A$ is quadratic in the gauge field, so that in $F_{\mu\nu}^{A2}$ we find cubic and quartic vertices beyond the kinetic term.

The QCD Lagrangian in eqn [2] has a simple structure but a very rich dynamical content, including the observed complex spectroscopy with a large number of hadrons. The most prominent properties of QCD are asymptotic freedom and confinement. In field theory, the effective coupling of a given interaction vertex is modified by the interaction. As a result, the measured intensity of the force depends on the transferred (four)momentum squared, Q^2 , among the participants. In QCD, the relevant coupling parameter that appears in physical processes is α_s (see eqn [4]). Asymptotic freedom means that the effective coupling becomes a function of

Q^2 : $\alpha_s(Q^2)$ decreases for increasing Q^2 and vanishes asymptotically. Thus, the QCD interaction becomes very weak in processes with large Q^2 , called hard processes or deep inelastic processes (i.e., with a final-state distribution of momenta and a particle content very different from that in the initial state). One can prove that in four spacetime dimensions all gauge theories based on a noncommuting group of symmetry are asymptotically free, and conversely. The effective coupling decreases very slowly at large momenta with the inverse logarithm of Q^2 : $\alpha_s(Q^2) = 1/b \log Q^2/\Lambda^2$, where b is a known constant and Λ is an energy of the order of a few hundred MeV. Since in quantum mechanics large momenta imply short wavelengths, the result is that at short distances the potential between two color charges is similar to the Coulomb potential, that is, proportional to $\alpha_s(r)/r$, with an effective color charge which is small at short distances. On the contrary the interaction strength becomes large at large distances or small transferred momenta, of order $Q \lesssim \Lambda$. In fact, the observed hadrons are tightly bound composite states of quarks, with compensating color charges so that they are overall neutral in color. The property of confinement is the impossibility of separating color charges, like individual quarks and gluons. This is because in QCD the interaction potential between color charges increases, at long distances, linearly in r . When we try to separate the quark and the antiquark that form a color-neutral meson the interaction energy grows until pairs of quarks and antiquarks are created from the vacuum and new neutral mesons are coalesced instead of free quarks. For example, consider the process $e^+e^- \rightarrow q\bar{q}$ at large center-of-mass energies. The final-state quark and antiquark have large energies, so they separate in opposite directions very fast. But the color-confinement forces create new pairs in between them. Two back-to-back jets of colorless hadrons are observed with a number of slow pions that make the exact separation of the two jets impossible. In some cases, a third well-separated jet of hadrons is also observed: these events correspond to the radiation of an energetic gluon from the parent quark-antiquark pair.

Electroweak Interactions

We split the electroweak Lagrangian into two parts by separating the Higgs boson couplings:

$$\mathcal{L} = \mathcal{L}_{\text{symm}} + \mathcal{L}_{\text{Higgs}} \quad [6]$$

We start by specifying $\mathcal{L}_{\text{symm}}$, which involves only gauge bosons and fermions (a sum over all flavors of

quarks and leptons, generally indicated by ψ is understood):

$$\mathcal{L}_{\text{symm}} = -\frac{1}{4} \sum_{A=1}^3 F_{\mu\nu}^A F^{A\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} + \bar{\psi}_L i \gamma^\mu D_\mu \psi_L + \bar{\psi}_R i \gamma^\mu D_\mu \psi_R \quad [7]$$

This is the Yang–Mills Lagrangian for the gauge group $SU(2) \otimes U(1)$ with fermion matter fields. Here

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$$

$$F_{\mu\nu}^A = \partial_\mu W_\nu^A - \partial_\nu W_\mu^A - g \epsilon_{ABC} W_\mu^B W_\nu^C \quad [8]$$

are the gauge antisymmetric tensors constructed out of the gauge field B_μ associated with $U(1)$, and W_μ^A corresponding to the three $SU(2)$ generators; ϵ_{ABC} are the group structure constants (see eqn [11]), which, for $SU(2)$, coincide with the totally antisymmetric Levi-Civita tensor (recall the familiar angular-momentum commutators).

The fermion fields are described through their left- and right-hand components:

$$\psi_{L,R} = [(1 \mp \gamma_5)/2] \psi, \quad \bar{\psi}_{L,R} = \bar{\psi} [(1 \pm \gamma_5)/2] \quad [9]$$

Note that, as given in eqn [9],

$$\bar{\psi}_L = \psi_L^\dagger \gamma_0 = \psi^\dagger [(1 - \gamma_5)/2] \gamma_0$$

$$= \bar{\psi} [\gamma_0 (1 - \gamma_5)/2] \gamma_0 = \bar{\psi} [(1 + \gamma_5)/2]$$

The matrices $P_\pm = (1 \pm \gamma_5)/2$ are projectors. They satisfy the relations $P_\pm P_\pm = P_\pm$, $P_\pm P_\mp = 0$, $P_+ + P_- = 1$.

The standard electroweak theory is a chiral theory, in the sense that ψ_L and ψ_R behave differently under the gauge group. In particular, all ψ_R are singlets and all ψ_L are doublets in the minimal SM (MSM). Thus, mass terms for fermions (of the form $\bar{\psi}_L \psi_R + \text{h.c.}$) are forbidden in the symmetric limit. Fermion masses are introduced, together with W^\pm and Z masses, by the mechanism of symmetry breaking. The covariant derivatives $D_\mu \psi_{L,R}$ are explicitly given by

$$D_\mu \psi_{L,R} = \left[\partial_\mu + ig \sum_{A=1}^3 t_{L,R}^A W_\mu^A + ig' \frac{1}{2} Y_{L,R} B_\mu \right] \psi_{L,R} \quad [10]$$

where $t_{L,R}^A$ and $1/2 Y_{L,R}$ are the $SU(2)$ and $U(1)$ generators, respectively, in the reducible representations $\psi_{L,R}$. The commutation relations of the $SU(2)$ generators are given by

$$[t_L^A, t_L^B] = i \epsilon_{ABC} t_L^C \quad \text{and} \quad [t_R^A, t_R^B] = i \epsilon_{ABC} t_R^C \quad [11]$$

We use the normalization $\text{tr}[t^A t^B] = 1/2 \delta^{AB}$ in the fundamental representation of $SU(2)$. The electric

charge generator Q (in units of e , the positron charge) is given by

$$Q = t_L^3 + 1/2 Y_L = t_R^3 + 1/2 Y_R \quad [12]$$

All fermion couplings to the gauge bosons can be derived directly from eqns [7] and [10]. The charged-current (CC) couplings are the simplest. From

$$\begin{aligned} g(t^1 W_\mu^1 + t^2 W_\mu^2) &= g \left\{ \left[(t^1 + i t^2)/\sqrt{2} \right] \right. \\ &\quad \times \left[(W_\mu^1 - i W_\mu^2)/\sqrt{2} \right] + \text{h.c.} \left. \right\} \\ &= g \left\{ \left[(t^+ W_\mu^-)/\sqrt{2} \right] + \text{h.c.} \right\} \quad [13] \end{aligned}$$

where $t^\pm = t^1 \pm i t^2$ and $W^\pm = (W^1 \pm i W^2)/\sqrt{2}$, we obtain the vertex

$$\begin{aligned} V_{\bar{\psi}\psi W} &= g \bar{\psi} \gamma_\mu \left[(t_L^+/\sqrt{2})(1 - \gamma_5)/2 + (t_R^+/\sqrt{2}) \right. \\ &\quad \times (1 + \gamma_5)/2 \left. \right] \psi W_\mu^- + \text{h.c.} \quad [14] \end{aligned}$$

In the neutral-current (NC) sector, the photon A_μ and the mediator Z_μ of the weak NC are orthogonal and normalized linear combinations of B_μ and W_μ^3 :

$$\begin{aligned} A_\mu &= \cos \theta_W B_\mu + \sin \theta_W W_\mu^3 \\ Z_\mu &= -\sin \theta_W B_\mu + \cos \theta_W W_\mu^3 \quad [15] \end{aligned}$$

Equations [15] define the weak mixing angle θ_W . The photon is characterized by equal couplings to left and right fermions with a strength equal to the electric charge. Recalling eqn [12] for the charge matrix Q , we immediately obtain

$$g \sin \theta_W = g' \cos \theta_W = e \quad [16]$$

or, equivalently,

$$\tan \theta_W = g'/g \quad [17]$$

Once θ_W has been fixed by the photon couplings, it is a simple matter of algebra to derive the Z couplings, with the result

$$\begin{aligned} \Gamma_{\bar{\psi}\psi Z} &= g/(2 \cos \theta_W) \bar{\psi} \gamma_\mu [t_L^3(1 - \gamma_5) + t_R^3(1 + \gamma_5) \\ &\quad - 2Q \sin^2 \theta_W] \psi Z^\mu \quad [18] \end{aligned}$$

where $\Gamma_{\bar{\psi}\psi Z}$ is a notation for the vertex. In the MSM, $t_R^3 = 0$ and $t_L^3 = \pm 1/2$. Note that the CC and NC weak couplings do not conserve P (parity) and C (charge conjugation).

In order to derive the effective four-fermion interactions that are equivalent, at low energies, to the CC and NC couplings given in eqns [14] and [18], we anticipate that large masses, as experimentally observed, are provided for W^\pm and Z by $\mathcal{L}_{\text{Higgs}}$. For left-left CC couplings, when the momentum

transfer squared can be neglected with respect to m_W^2 in the propagator of Born diagrams with single W exchange, from eqn [14], we can write

$$\begin{aligned} \mathcal{L}_{\text{eff}}^{\text{CC}} &\simeq (g^2/8m_W^2) [\bar{\psi} \gamma_\mu (1 - \gamma_5) t_L^+ \psi] \\ &\quad \times [\bar{\psi} \gamma^\mu (1 - \gamma_5) t_L^- \psi] \quad [19] \end{aligned}$$

By specializing further in the case of doublet fields such as $\nu_e - e^-$ or $\nu_\mu - \mu^-$, we obtain the tree-level relation of g with the Fermi coupling constant G_F measured from μ decay ($G_F = 1.16639(2) \times 10^{-5} \text{ GeV}^{-2}$):

$$G_F/\sqrt{2} = g^2/8m_W^2 \quad [20]$$

By recalling that $g \sin \theta_W = e$, we can also cast this relation in the form

$$m_W = \mu_{\text{Born}}/\sin \theta_W \quad [21]$$

with

$$\mu_{\text{Born}} = (\pi\alpha/\sqrt{2}G_F)^{1/2} \simeq 37.2802 \text{ GeV} \quad [22]$$

where α is the fine-structure constant of QED ($\alpha \equiv e^2/4\pi = 1/137.036$).

In the same way, for neutral currents we obtain, in Born approximation, from eqn [18], the effective four-fermion interaction given by

$$\mathcal{L}_{\text{eff}}^{\text{NC}} \simeq \sqrt{2} G_F \rho_0 \bar{\psi} \gamma_\mu [\dots] \psi \bar{\psi} \gamma^\mu [\dots] \psi \quad [23]$$

where

$$[\dots] \equiv t_L^3(1 - \gamma_5) + t_R^3(1 + \gamma_5) - 2Q \sin^2 \theta_W \quad [24]$$

and

$$\rho_0 = m_W^2/m_Z^2 \cos^2 \theta_W \quad [25]$$

All couplings given in this section are obtained at tree level and are modified in higher orders of perturbation theory. In particular, the relations between m_W and $\sin \theta_W$ (eqns [21] and [22]) and the observed values of ρ ($\rho = \rho_0$ at tree level) in different NC processes are altered by computable small electroweak radiative corrections.

The gauge-boson self-interactions can be derived from the $F_{\mu\nu}$ term in $\mathcal{L}_{\text{symm}}$, by using eqn [15] and $W^\pm = (W^1 \pm i W^2)/\sqrt{2}$. For the three-gauge-boson vertex $W^+ W^- V$ with $V = Z, \gamma$, we obtain

$$\begin{aligned} \Gamma_{W^- W^+ V} &= i g_{W^- W^+ V} [g_{\mu\nu}(q - p)_\lambda + g_{\mu\lambda}(p - r)_\nu \\ &\quad + g_{\nu\lambda}(r - q)_\mu] \quad [26] \end{aligned}$$

with

$$\begin{aligned} g_{W^- W^+ \gamma} &= g \sin \theta_W = e \quad \text{and} \\ g_{W^- W^+ Z} &= g \cos \theta_W \quad [27] \end{aligned}$$

This form of the triple gauge vertex is very special: in general, there could be departures from the above SM expression, even restricting us to $SU(2) \otimes U(1)$ gauge symmetric and C and P invariant couplings. In fact, some small corrections are already induced by the radiative corrections. The SM form of the triple gauge vertex has been experimentally confirmed by measuring the cross section $e^+e^- \rightarrow W^+W^-$ at LEP.

We now turn to the Higgs sector of the electroweak Lagrangian. The Higgs Lagrangian is specified by the gauge principle and the requirement of renormalizability to be

$$\mathcal{L}_{\text{Higgs}} = (D_\mu \phi)^\dagger (D^\mu \phi) - V(\phi^\dagger \phi) - \bar{\psi}_L \Gamma \psi_R \phi - \bar{\psi}_R \Gamma^\dagger \psi_L \phi^\dagger \quad [28]$$

where ϕ is a column vector including all Higgs scalar fields; it transforms as a reducible representation of the gauge group. The quantities Γ (which include all coupling constants) are matrices that make the Yukawa couplings invariant under the Lorentz and gauge groups. The potential $V(\phi^\dagger \phi)$, symmetric under $SU(2) \otimes U(1)$, contains, at most, quartic terms in ϕ so that the theory is renormalizable:

$$V(\phi^\dagger \phi) = -\frac{1}{2}\mu^2 \phi^\dagger \phi + \frac{1}{4}\lambda(\phi^\dagger \phi)^2 \quad [29]$$

Spontaneous symmetry breaking is induced if the minimum of V , which is the classical analog of the quantum-mechanical vacuum state (both are the states of minimum energy) is obtained for nonvanishing ϕ values. This occurs because we have taken μ^2 and λ positive in V (note the “wrong” sign of the mass term). Precisely, we denote the vacuum expectation value (VEV) of ϕ , that is, the position of the minimum, by v :

$$\langle 0|\phi(x)|0\rangle = v \neq 0 \quad [30]$$

The fermion mass matrix is obtained from the Yukawa couplings by replacing $\phi(x)$ by v :

$$M = \bar{\psi}_L \mathcal{M} \psi_R + \bar{\psi}_R \mathcal{M}^\dagger \psi_L \quad [31]$$

with

$$\mathcal{M} = \Gamma \cdot v \quad [32]$$

In the SM, where all left fermions, ψ_L , are doublets and all right fermions, ψ_R , are singlets, only Higgs doublets can contribute to fermion masses. There are enough free couplings in Γ , so that one single complex Higgs doublet is indeed sufficient to generate the most general fermion mass matrix. It is important to observe that by a suitable change of basis we can always make the matrix \mathcal{M} Hermitian,

γ_5 -free and diagonal. In fact, we can make separate unitary transformations on ψ_L and ψ_R according to

$$\psi'_L = U\psi_L, \quad \psi'_R = V\psi_R \quad [33]$$

and consequently

$$\mathcal{M} \rightarrow \mathcal{M}' = U^\dagger \mathcal{M} V \quad [34]$$

This transformation does not alter the general structure of the fermion couplings in \mathcal{L}_{sym} .

If only one Higgs doublet is present, the change of basis that makes \mathcal{M} diagonal will at the same time diagonalize also the fermion–Higgs Yukawa couplings. Thus, in this case, no flavor-changing neutral Higgs exchanges are present. This is not true, in general, when there are several Higgs doublets. But one Higgs doublet for each electric charge sector, that is, one doublet coupled only to u -type quarks, one doublet to d -type quarks, one doublet to charged leptons would also be satisfactory, because the mass matrices of fermions with different charges are diagonalized separately. In fact, at the moment, the simplest model with only one Higgs doublet seems adequate for describing all observed phenomena.

Weak charged currents are the only tree-level interactions in the SM that change flavor: by emission of a W , a u -type quark is turned into a d -type quark, or a ν_l neutrino is turned into an l^- charged lepton (all fermions are left-handed). If we start from a u -type quark that is a mass eigenstate, emission of a W turns it into a d -type quark state d' (the weak isospin partner of u) that in general is not a mass eigenstate. In general, the mass eigenstates and the weak eigenstates do not coincide and a unitary transformation connects the two sets:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V \begin{pmatrix} d \\ s \\ b \end{pmatrix} \quad [35]$$

or, in shorthand, $D' = VD$, where V is the Cabibbo–Kobayashi–Maskawa (CKM) matrix. Thus, in terms of mass eigenstates the charged weak current of quarks is of the form

$$J_\mu^+ \propto \bar{u} \gamma_\mu (1 - \gamma_5) VD \quad [36]$$

Since V is unitary (i.e., $VV^\dagger = V^\dagger V = 1$) and commutes with T^2 , T_3 , and Q (because all d -type quarks have the same isospin and charge) the neutral current couplings are diagonal both in the primed and unprimed basis (if the Z d -type quark current is abbreviated as $\bar{D}'\Gamma D'$ then by changing basis we get $\bar{D}V^\dagger\Gamma V D$ and V and Γ commute because, as seen from eqn [24], Γ is made of Dirac matrices and T_3 and Q generator matrices). It follows that $\bar{D}'\Gamma D' = \bar{D}\Gamma D$. This is the Glashow–Iliopoulos–Maiani (GIM)

mechanism that ensures natural flavor conservation of the neutral current couplings at the tree level. For three generations of quarks, the CKM matrix depends on four physical parameters: three mixing angles and one phase. This phase is the unique source of CP violation in the SM.

We now consider the gauge-boson masses and their couplings to the Higgs. These effects are induced by the $(D_\mu\phi)^\dagger(D^\mu\phi)$ term in $\mathcal{L}_{\text{Higgs}}$ (eqn [28]), where

$$D_\mu\phi = \left[\partial_\mu + ig \sum_{A=1}^3 t^A W_\mu^A + ig'(Y/2)B_\mu \right] \phi \quad [37]$$

Here t^A and $1/2Y$ are the $SU(2) \otimes U(1)$ generators in the reducible representation spanned by ϕ . Not only doublets but all non-singlet Higgs representations can contribute to gauge-boson masses. The condition that the photon remains massless is equivalent to the condition that the vacuum is electrically neutral:

$$Q|\nu\rangle = (t^3 + \tfrac{1}{2}Y)|\nu\rangle = 0 \quad [38]$$

The charged W mass is given by the quadratic terms in the W field arising from $\mathcal{L}_{\text{Higgs}}$, when $\phi(x)$ is replaced by ν . We obtain

$$m_W^2 W_\mu^+ W^{-\mu} = g^2 \left| \left(t^+ \nu / \sqrt{2} \right) \right|^2 W_\mu^+ W^{-\mu} \quad [39]$$

whilst for the Z mass we get (recalling eqn [15])

$$\begin{aligned} \tfrac{1}{2} m_Z^2 Z_\mu Z^\mu = & \left| [g \cos \theta_W t^3 \right. \\ & \left. - g' \sin \theta_W (Y/2)] \nu \right|^2 Z_\mu Z^\mu \end{aligned} \quad [40]$$

where the factor of $1/2$ on the left-hand side is the correct normalization for the definition of the mass of a neutral field. For Higgs doublets

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}, \quad \nu = \begin{pmatrix} 0 \\ \nu \end{pmatrix} \quad [41]$$

we obtain

$$m_W^2 = 1/2 g^2 \nu^2, \quad m_Z^2 = 1/2 g^2 \nu^2 / \cos^2 \theta_W \quad [42]$$

Note that by using eqn [20] we obtain

$$\nu = 2^{-3/4} G_F^{-1/2} = 174.1 \text{ GeV} \quad [43]$$

It is also evident that for Higgs doublets

$$\rho_0 = m_W^2 / m_Z^2 \cos^2 \theta_W = 1 \quad [44]$$

This relation is typical of one or more Higgs doublets and would be spoiled by the existence of, for example, Higgs triplets. This result is valid at the tree level and is modified by calculable small electroweak radiative corrections. The ρ_0 parameter has been measured from the intensity of NC interactions (recall eqn [25]) and confirmed to be close to unity at a few per milli level.

In MSM only one Higgs doublet is present. Then the fermion–Higgs couplings are in proportion to the fermion masses. In fact, from the Yukawa couplings $g_{\phi\bar{f}f}(\bar{f}_L\phi f_R + \text{h.c.})$, the mass m_f is obtained by replacing ϕ by ν , so that $m_f = g_{\phi\bar{f}f}\nu$. In MSM, three out of the four Hermitian fields are removed from the physical spectrum by the Higgs mechanism and become the longitudinal modes of W^+ , W^- , and Z which acquire a mass. The fourth neutral Higgs is physical and should be found. If more doublets are present, two more charged and two more neutral Higgs scalars should be around for each additional doublet.

The couplings of the physical Higgs H to the gauge bosons can be simply obtained from $\mathcal{L}_{\text{Higgs}}$, by the replacement

$$\phi(x) = \begin{pmatrix} \phi^+(x) \\ \phi^0(x) \end{pmatrix} \rightarrow \begin{pmatrix} 0 \\ \nu + (H/\sqrt{2}) \end{pmatrix} \quad [45]$$

(so that $(D_\mu\phi)^\dagger(D^\mu\phi) = 1/2(\partial_\mu H)^2 + \dots$), with the result

$$\begin{aligned} \mathcal{L}[H, W, Z] &= g^2 (\nu/\sqrt{2}) W_\mu^+ W^{-\mu} H + (g^2/4) W_\mu^+ W^{-\mu} H^2 \\ &+ \left[(g^2 \nu Z_\mu Z^\mu) / (2\sqrt{2} \cos^2 \theta_W) \right] H \\ &+ [g^2 / (8 \cos^2 \theta_W)] Z_\mu Z^\mu H^2 \end{aligned}$$

In MSM, the Higgs mass $m_H^2 \sim \lambda \nu^2$ is of order of the weak scale ν but cannot be predicted because the value of λ is not fixed. The dominant decay mode of the Higgs is in the $b\bar{b}$ channel below the WW threshold, while the W^+W^- channel is dominant for sufficiently large m_H . The width is small below the WW threshold, not exceeding a few MeV, but increases steeply beyond the threshold, reaching the asymptotic value of $\Gamma \sim 1/2 m_H^3$ at large m_H , where all energies and masses are in TeV.

A central role in the experimental verification of the standard electroweak theory has been played by CERN, the European Laboratory for Particle Physics, located near Geneva, between France and Switzerland. The indirect effects of the Z_0 , that is, the occurrence of weak processes induced by the neutral current, were first observed in 1974 at CERN by the Collaboration Gargamelle (the name of the bubble chamber used in the experiment). Later, in 1982, the W^\pm and the Z_0 were, for the first time, directly produced and observed in proton–antiproton collisions by the UA1 and UA2 collaborations and then further studied with the same technique both at CERN and subsequently at the Tevatron of Fermilab near Chicago. Starting from 1989 LEP, the large e^+e^- collider was functioning at CERN till 2000. In the LEP circular ring of circumference ~ 27 km, electrons and

positrons were accelerated in opposite directions to an equal energy in the range between 45 and 103 GeV. The beams were made to cross and collide in correspondence of four experimental areas where the ALEPH, DELPHI, L3, and OPAL detectors were located to study the final states produced in the collisions. In its first phase, called LEP1, from 1989 to 1995 the LEP operation had been completely dedicated to a precise study of the Z_0 properties, mass, lifetime, and decay modes in order to accurately test the predictions of the SM. The main lessons of the precision tests of the standard electroweak theory can be summarized as follows. It has been checked that the couplings of quarks and leptons to the weak gauge bosons W^\pm and Z are indeed precisely those prescribed by the gauge symmetry. The accuracy of a few tenths of 1% for these tests implies that, not only the tree level, but also the structure of quantum corrections has been verified. Then, since the end of 1995, the energy of LEP was increased and the phase of LEP2 was started. The total energy was gradually increased up to 206 GeV. The main physics goals of LEP2 were the search for the Higgs and for possible new particles, the precise measurement of m_W and the experimental study of the triple gauge vertices $WW\gamma$ and WWZ_0 . The Higgs particle of the SM could in principle be produced at LEP2 in the reaction $e + e^- \rightarrow Z_0 H$, which proceeds by Z_0 exchange. The nonobservation of the Higgs particle at LEP2 has allowed to establish a lower limit on its mass: $m_H \gtrsim 114$ GeV. Indirect indications on the Higgs mass were also obtained from the precision tests of the SM, as the radiative effects depend logarithmically on m_H . The indication is that the Higgs mass cannot be too heavy if the SM is valid: $m_H \lesssim 219$ GeV at 95% c.l. In 2001, LEP was

dismantled and, in its tunnel, a new double ring of superconducting magnets is being installed. The new accelerator, the LHC (Large Hadron Collider), will be a proton–proton collider of total center-of-mass energy 14 TeV. Two large experiments ATLAS and CMS will continue to search for the Higgs starting in the year 2007. The sensitivity of LHC experiments to the SM Higgs will go up to masses m_H of ~ 1 TeV.

See also: Effective Field Theories; Electric–Magnetic Duality; Electroweak Theory; General Relativity: Experimental Tests; Noncommutative Geometry and the Standard Model; Perturbative Renormalization Theory and BRST; Quantum Chromodynamics; Quantum Electrodynamics and its Precision Tests; Quantum Field Theory: a Brief Introduction; Relativistic Wave Equations Including Higher Spin Fields; Renormalization: General Theory; Supersymmetric Particle Models.

Further Reading

- Altarelli G (2000) *The Standard Electroweak Theory and Beyond*, Proceedings of the Summer School on Phenomenology of Gauge Interactions. Zuoz, Switzerland hep-ph/0011078.
- Altarelli G (2001) *A QCD Primer*, Proceedings of the 2001 European School of High Energy Physics. Beatenberg, Switzerland, hep-ph/0204179.
- Close F, Marten, and Sutton MC (2002) *The Particle Odyssey*. Oxford: Oxford University Press.
- Fraser G (ed.) (1998) *The Particle Century*. London: The Institute of Physics.
- Martin BR and Shaw G (1997) *Particle Physics*, 2nd edn. Chichester: Wiley.
- Particle Data Group (2004) Review of particle physics. *Physics Letters B* 592: 1.
- Perkins DH (2000) *Introduction to High Energy Physics*. Reading: Addison-Wesley.

Stationary Black Holes

R Beig, Universität Wien, Vienna, Austria

P T Chruściel, Université de Tours, Tours, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction

This article treats a specific class of stationary solutions to the Einstein field equations which read

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \frac{8\pi G}{c^4} T_{\mu\nu} \quad [1]$$

Here $R_{\mu\nu}$ and $R = g^{\mu\nu} R_{\mu\nu}$ are, respectively, the Ricci tensor and the Ricci scalar of the spacetime metric $g_{\mu\nu}$, G the Newton constant, and c the speed of light.

The tensor $T_{\mu\nu}$ is the stress–energy tensor of matter. Spacetimes, or regions thereof, where $T_{\mu\nu} = 0$ are called vacuum.

Stationary solutions are of interest for a variety of reasons. As models for compact objects at rest, or in steady rotation, they play a key role in astrophysics. They are easier to study than nonstationary systems because stationary solutions are governed by elliptic rather than hyperbolic equations. Finally, like in any field theory, one expects that large classes of dynamical solutions approach (“settle down to”) a stationary state in the final stages of their evolution.

The simplest stationary solutions describing compact isolated objects are the spherically symmetric

ones. In the vacuum region, these are all given by the Schwarzschild family. A theorem of Birkhoff shows that in the vacuum region any spherically symmetric metric, even without assuming stationarity, belongs to the family of Schwarzschild metrics, parametrized by a positive mass parameter m . Thus, regardless of possible motions of the matter, as long as they remain spherically symmetric, the exterior metric is the Schwarzschild one for some constant m . This has the following consequence for stellar dynamics: imagine following the collapse of a cloud of pressureless fluid (“dust”). Within Newtonian gravity, this dust cloud will, after finite time, contract to a point at which the density and the gravitational potential diverge. However, this result cannot be trusted as a sensible physical prediction because, even if one supposes that Newtonian gravity is still valid at very high densities, a matter model based on noninteracting point particles is certainly not. Consider, next, the same situation in the Einstein theory of gravity: here a new question arises, related to the form of the Schwarzschild metric outside of the spherically symmetric body:

$$\begin{aligned} g &= -V^2 dt^2 + V^{-2} dr^2 + r^2 d\Omega^2, \\ V^2 &= 1 - \frac{2Gm}{rc^2}, \\ t &\in \mathbb{R}, \quad r \in \left(\frac{2Gm}{c^2}, \infty\right) \end{aligned} \quad [2]$$

Here $d\Omega^2$ is the line element of the standard 2-sphere. Since the metric [2] seems to be singular as $r = 2m$ is approached (from now on, we use units in which $G = c = 1$), there arises the need to understand what happens at the surface of the star when the radius $r = 2m$ is reached. One thus faces the need of a careful study of the geometry of the metric [2] when $r = 2m$ is approached, and crossed.

The first key feature of the metric [2] is its stationarity, of course, with Killing vector field X given by $X = \partial_t$. A Killing field, by definition, is a vector field the local flow of which generates isometries. A spacetime (the term spacetime denotes a smooth, paracompact, connected, orientable, and time-orientable Lorentzian manifold) is called stationary if there exists a Killing vector field X which approaches ∂_t in the asymptotically flat region (where r goes to ∞ ; see below for precise definitions) and generates a one-parameter group of isometries. A spacetime is called static if it is stationary and if the stationary Killing vector X is hypersurface orthogonal, that is, $X^\flat \wedge dX^\flat = 0$, where $X^\flat = X_\mu dx^\mu = g_{\mu\nu} X^\nu dx^\mu$. A spacetime is called axisymmetric if there exists a Killing vector field Y , which generates a one-parameter group of isometries and which behaves like a rotation

in the asymptotically flat region, with all orbits 2π -periodic. In asymptotically flat spacetimes, this implies that there exists an axis of symmetry, that is, a set on which the Killing vector vanishes. Killing vector fields which are a nontrivial linear combination of a time translation and of a rotation in the asymptotically flat region are called stationary rotating, or helical.

There exists a technique, due independently to Kruskal and Szekeres, of attaching together two regions $r > 2m$ and two regions $r < 2m$ of the Schwarzschild metric, as in Figure 1, to obtain a manifold with a metric which is smooth at $r = 2m$. In the extended spacetime, the hypersurface $\{r = 2m\}$ is a null hypersurface \mathcal{E} , the Schwarzschild event horizon. The stationary Killing vector $X = \partial_t$ extends to a Killing vector in the extended spacetime which becomes tangent to and null on \mathcal{E} . The global properties of the Kruskal–Szekeres extension of the exterior Schwarzschild spacetime make this spacetime a natural model for a nonrotating black hole. It is worth noting here that the exterior Schwarzschild spacetime [2] admits an infinite number of nonisometric vacuum extensions, even in the class of maximal, analytic, simply connected ones. The Kruskal–Szekeres extension is singled out by the properties that it is maximal, vacuum, analytic, simply connected, with all maximally extended geodesics either complete, or with the area r of the orbits of the isometry groups tending to zero along them.

We can now come back to the problem of the contracting dust cloud according to the Einstein theory. For simplicity, we take the density of the dust to be uniform – the so-called Oppenheimer–Snyder solution. It then turns out that, in the course of collapse, the surface of the dust will eventually cross the Schwarzschild radius, leaving behind a Schwarzschild black hole. If one follows the dust cloud further, a singularity will eventually form, but will not be visible from the “outside region” where $r > 2m$. For a collapsing body of the mass of the Sun, say, one has $2m = 3$ km. Thus, standard phenomenological matter models such as that for dust can still be trusted, so that the previous objection to the Newtonian scenario does not apply.

There is a rotating generalization of the Schwarzschild metric, namely the two-parameter family of exterior Kerr metrics, which in Boyer–Lindquist coordinates takes the form

$$\begin{aligned} g &= -\frac{\Delta - a^2 \sin^2 \theta}{\Sigma} dt^2 - \frac{2a \sin^2 \theta (r^2 + a^2 - \Delta)}{\Sigma} dt d\varphi \\ &\quad + \frac{(r^2 + a^2)^2 - \Delta a^2 \sin^2 \theta}{\Sigma} \sin^2 \theta d\varphi^2 \\ &\quad + \frac{\Sigma}{\Delta} dr^2 + \Sigma d\theta^2 \end{aligned} \quad [3]$$

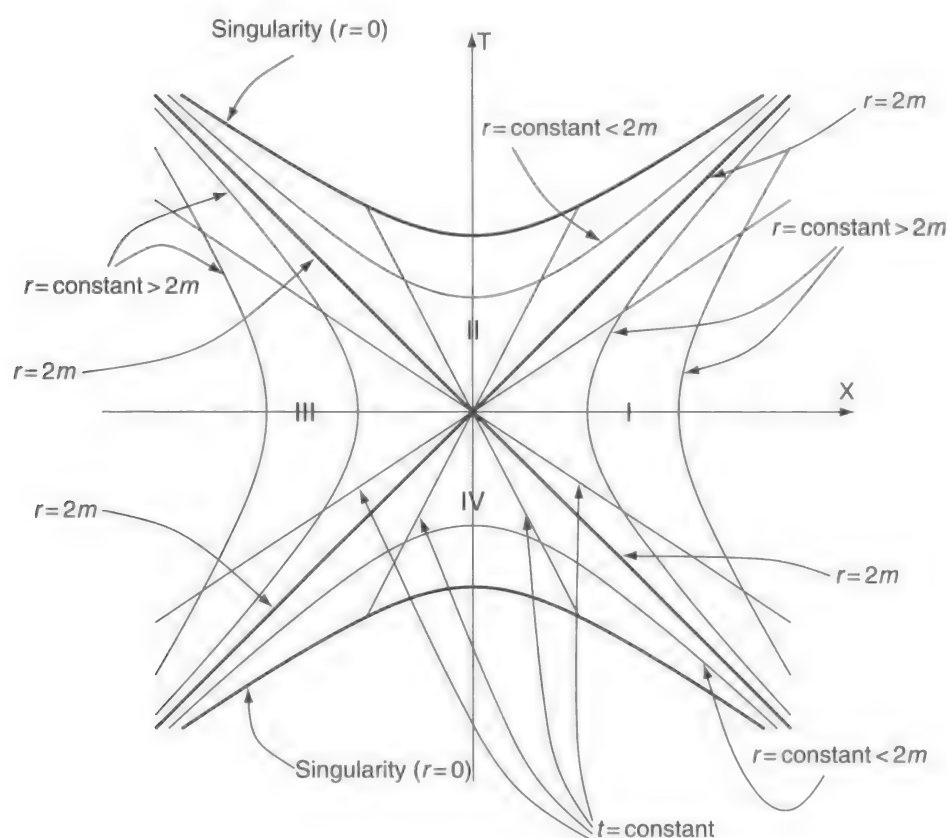


Figure 1 The Kruskal–Szekeres extension of the Schwarzschild solution. (Adapted with permission from Nicolas J-P (2002) Dirac fields on asymptotically flat space-times. *Dissertationes Mathematicae* 408: 1–85.)

with $0 \leq a < m$. Here $\Sigma = r^2 + a^2 \cos^2 \theta$, $\Delta = r^2 + a^2 - 2mr$ and $r_+ < r < \infty$ where $r_+ = m + (m^2 - a^2)^{1/2}$. When $a=0$, the Kerr metric reduces to the Schwarzschild metric. The Kerr metric is again a vacuum solution, and it is stationary with $X = \partial_t$ the asymptotic time translation, as well as axisymmetric with $Y = \partial_\phi$ the generator of rotations. Similarly to the Schwarzschild case, it turns out that the metric can be smoothly extended across $r=r_+$, with $\{r=r_+\}$ being a smooth null hypersurface \mathcal{E} in the extension. The null generator K of \mathcal{E} is the limit of the stationary-rotating Killing field $X + \omega Y$, where $\omega = a/(2mr_+)$. On the other hand, the Killing vector X is timelike only outside the hypersurface $\{r = m + (m^2 - a^2 \cos^2 \theta)^{1/2}\}$, on which X becomes null. In the region between r_+ and $r = m + (m^2 - a^2 \cos^2 \theta)^{1/2}$, which is called the ergoregion, X is spacelike. It is also spacelike on and tangent to \mathcal{E} , except where the axis of rotation meets \mathcal{E} , where X is null. Based on the above properties, the Kerr family provides natural models for rotating black holes.

Unfortunately, as opposed to the spherically symmetric case, there are no known explicit collapsing solutions with rotating matter, in particular no known solutions having the Kerr metric as final state.

The aim of the theory outlined below is to understand the general geometrical features of

stationary black holes, and to give a classification of models satisfying the field equations.

Model-Independent Concepts

Some of the notions used informally in the introductory section will now be made more precise. The mathematical notion of black hole is meant to capture the idea of a region of spacetime which cannot be seen by “outside observers.” Thus, at the outset, one assumes that there exists a family of physically preferred observers in the spacetime under consideration. When considering isolated physical systems, it is natural to define the “exterior observers” as observers which are “very far” away from the system under consideration. The standard way of making this mathematically precise is by using conformal completions, discussed in more detail in the article about asymptotic structure in this encyclopedia: a pair $(\tilde{\mathcal{M}}, \tilde{g})$ is called a conformal completion at infinity, or simply conformal completion, of (\mathcal{M}, g) if $\tilde{\mathcal{M}}$ is a manifold with boundary such that:

1. \mathcal{M} is the interior of $\tilde{\mathcal{M}}$;
2. there exists a function Ω , with the property that the metric \tilde{g} , defined as $\Omega^2 g$ on \mathcal{M} , extends by continuity to the boundary of $\tilde{\mathcal{M}}$, with the

extended metric remaining of Lorentzian signature; and

3. Ω is positive on \mathcal{M} , differentiable on $\tilde{\mathcal{M}}$, vanishes on the boundary

$$\mathcal{I} := \tilde{\mathcal{M}} \setminus \mathcal{M}$$

with $d\Omega$ nowhere vanishing on \mathcal{I} .

The boundary \mathcal{I} of $\tilde{\mathcal{M}}$ is called Scri, a phonic shortcut for “script I.” The idea here is the following: forcing Ω to vanish on \mathcal{I} ensures that \mathcal{I} lies infinitely far away from any physical object – a mathematical way of capturing the notion “very far away.” The condition that $d\Omega$ does not vanish is a convenient technical condition which ensures that \mathcal{I} is a smooth three-dimensional hypersurface, instead of some, say, one- or two-dimensional object, or of a set with singularities here and there. Thus, \mathcal{I} is an idealized description of a family of observers at infinity.

To distinguish between various points of \mathcal{I} , one sets

$$\mathcal{I}^+ = \{\text{points in } \mathcal{I} \text{ which are to the future of the physical spacetime}\}$$

$$\mathcal{I}^- = \{\text{points in } \mathcal{I} \text{ which are to the past of the physical spacetime}\}$$

(Recall that a point q is to the future, respectively to the past, of p if there exists a future directed, respectively past directed, causal curve from p to q . Causal curves are curves γ such that their tangent vector $\dot{\gamma}$ is causal everywhere, $g(\dot{\gamma}, \dot{\gamma}) \leq 0$.) One then defines the black hole region \mathcal{B} as

$$\mathcal{B} := \{\text{points in } \mathcal{M} \text{ which are not in the past of } \mathcal{I}^+\} \quad [4]$$

By definition, points in the black hole region cannot thus send information to \mathcal{I}^+ ; equivalently, observers on \mathcal{I}^+ cannot see points in \mathcal{B} . The white-hole region \mathcal{W} is defined by changing the time orientation in [4]. A key notion related to the concept of a black hole is that of future (\mathcal{E}^+) and past (\mathcal{E}^-) event horizons,

$$\mathcal{E}^+ := \partial\mathcal{B}, \quad \mathcal{E}^- := \partial\mathcal{W} \quad [5]$$

Under mild assumptions, event horizons in stationary spacetimes with matter satisfying the null-energy condition,

$$T_{\mu\nu}\ell^\mu\ell^\nu \geq 0 \quad \text{for all null vectors } \ell^\mu \quad [6]$$

are smooth null hypersurfaces, analytic if the metric is analytic.

In order to develop a reasonable theory, one also needs a regularity condition for the interior of

spacetime. This has to be a condition which does not exclude singularities (otherwise the Schwarzschild and Kerr black holes would be excluded), but which nevertheless guarantees a well-behaved exterior region. One such condition, assumed in all the results described below, is the existence in \mathcal{M} of an asymptotically flat spacelike hypersurface \mathcal{S} with compact interior. Further, either \mathcal{S} has no boundary or the boundary of \mathcal{S} lies on $\mathcal{E}^+ \cup \mathcal{E}^-$. To make things precise, for any spacelike hypersurface let g_{ij} be the induced metric, and let K_{ij} denote its extrinsic curvature. A spacelike hypersurface \mathcal{S}_{ext} diffeomorphic to \mathbb{R}^3 minus a ball will be called asymptotically flat if the fields (g_{ij}, K_{ij}) satisfy the fall-off conditions

$$|g_{ij} - \delta_{ij}| + r|\partial_i g_{ij}| + \cdots + r^k|\partial_{\ell_1 \dots \ell_k} g_{ij}| + r|K_{ij}| + \cdots + r^k|\partial_{\ell_1 \dots \ell_k} K_{ij}| \leq Cr^{-1} \quad [7]$$

for some constants $C, k \geq 1$. A hypersurface \mathcal{S} (with or without boundary) will be said to be asymptotically flat with compact interior if \mathcal{S} is of the form $\mathcal{S}_{\text{int}} \cup \mathcal{S}_{\text{ext}}$, with \mathcal{S}_{int} compact and \mathcal{S}_{ext} asymptotically flat.

There exists a canonical way of constructing a conformal completion with good global properties for stationary spacetimes which are asymptotically flat in the sense of [7], and which are vacuum sufficiently far out in the asymptotic region. This conformal completion is referred to as the standard completion and will be assumed from now on.

Returning to the event horizon $\mathcal{E} = \mathcal{E}^+ \cup \mathcal{E}^-$, it is not very difficult to show that every Killing vector field X is necessarily tangent to \mathcal{E} . Since the latter set is a null Lipschitz hypersurface, it follows that X is either null or spacelike on \mathcal{E} . This leads to a preferred class of event horizons, called Killing horizons. By definition, a Killing horizon associated with a Killing vector K is a null hypersurface which coincides with a connected component of the set

$$\mathcal{H}(K) := \{p \in \mathcal{M} : g(K, K)(p) = 0, K(p) \neq 0\} \quad [8]$$

A simple example is provided by the “boost Killing vector field” $K = z\partial_t + t\partial_z$ in Minkowski spacetime: $\mathcal{H}(K)$ has four connected components,

$$\mathcal{H}_{\epsilon\delta} := \{t = \epsilon z, \delta t > 0\}, \quad \epsilon, \delta \in \{\pm 1\}$$

The closure $\bar{\mathcal{H}}$ of \mathcal{H} is the set $\{|t| = |z|\}$, which is not a manifold, because of the crossing of the null hyperplanes $\{t = \pm z\}$ at $t = z = 0$. Horizons of this type are referred to as bifurcate Killing horizons, with the set $\{K(p) = 0\}$ being called the bifurcation surface of $\mathcal{H}(K)$. The bifurcate horizon structure in the Kruskal–Szekeres–Schwarzschild spacetime can be clearly seen in Figures 1 and 2.

The Vishveshwara–Carter lemma shows that if a Killing vector K is hypersurface orthogonal, $K^\flat \wedge dK^\flat = 0$, then the set $\mathcal{H}(K)$ defined in [8] is a union of smooth null hypersurfaces, with K being tangent to the null geodesics threading \mathcal{H} (“ \mathcal{H} is generated by K ”), and so is indeed a Killing horizon. It has been shown by Carter that the same conclusion can be reached if the hypothesis of hypersurface orthogonality is replaced by that of existence of two linearly independent Killing vector fields.

In stationary-axisymmetric spacetimes, a Killing vector K tangent to the generators of a Killing horizon \mathcal{H} can be normalized so that $K = X + \omega Y$, where X is the Killing vector field which asymptotes to a time translation in the asymptotic region, and Y is the Killing vector field which generates rotations in the asymptotic region. The constant ω is called the angular velocity of the Killing horizon \mathcal{H} .

On a Killing horizon $\mathcal{H}(K)$, one necessarily has

$$\nabla^\mu (K^\nu K_\nu) = -2\kappa K^\mu \quad [9]$$

Assuming the so-called dominant-energy condition on $T_{\mu\nu}$ (see Positive Energy Theorem and Other Inequalities in GR), it can be shown that κ is constant (recall that Killing horizons are always connected in the terminology used in this article); it is called the surface gravity of \mathcal{H} . A Killing horizon is called degenerate when $\kappa = 0$, and nondegenerate otherwise; by an abuse of terminology, one similarly talks of degenerate black holes, etc. In Kerr spacetimes we

have $\kappa = 0$ if and only if $m = a$. A fundamental theorem of Boyer shows that degenerate horizons are closed. This implies that a horizon $\mathcal{H}(K)$ such that K has zeros in $\bar{\mathcal{H}}$ is nondegenerate, and is of bifurcate type, as described above. Further, a nondegenerate Killing horizon with complete geodesic generators always contains zeros of K in its closure. However, it is not true that existence of a nondegenerate horizon implies that of zeros of K : take the Killing vector field $z\partial_t + t\partial_z$ in Minkowski spacetime from which the 2-plane $\{z = t = 0\}$ has been removed. The universal cover of that last spacetime provides a spacetime in which one cannot restore the points which have been artificially removed, without violating the manifold property.

The domain of outer communications (DOC) of a black hole spacetime is defined as

$$\langle\langle \mathcal{M} \rangle\rangle := \mathcal{M} \setminus \{\mathcal{B} \cup \mathcal{W}\} \quad [10]$$

Thus, $\langle\langle \mathcal{M} \rangle\rangle$ is the region lying outside of the white-hole region and outside of the black hole region; it is the region which can both be seen by the outside observers and influenced by those.

The subset of $\langle\langle \mathcal{M} \rangle\rangle$ where X is spacelike is called the ergoregion. In the Schwarzschild spacetime, $\omega = 0$ and the ergoregion is empty, but neither of these is true in Kerr with $a \neq 0$.

A very convenient method for visualizing the global structure of spacetimes is provided by the Carter–Penrose diagrams. An example of such a diagram is given in Figure 2.

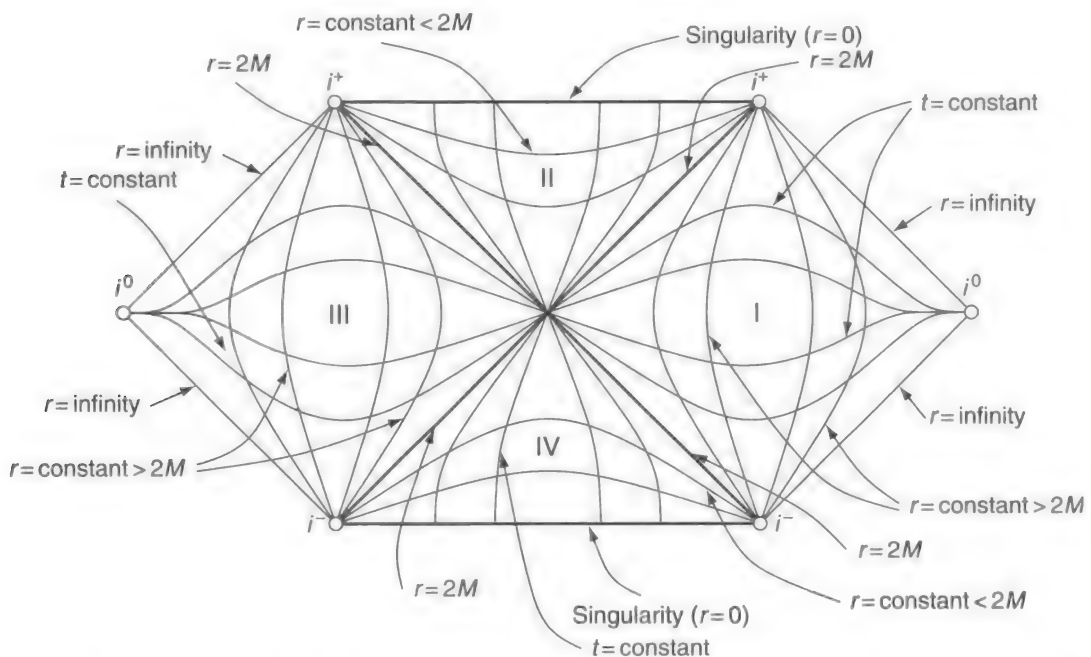


Figure 2 The Carter–Penrose diagram for the Kruskal–Szekeres spacetime. There are actually two asymptotically flat regions, with corresponding \mathcal{I}^\pm and \mathcal{S}^\pm defined with respect to the second region, but not indicated on this diagram. Each point in this diagram represents a two-dimensional sphere, and coordinates are chosen so that light cones have slopes ± 1 . Regions are numbered as in Figure 1. (Adapted with permission from Nicolas J-P (2002) Dirac fields on asymptotically flat space-times. *Dissertationes Mathematicae* 408: 1–85.)

A corollary of the topological censorship theorem of Friedman, Schleich, and Witt is that DOCs of regular black hole spacetimes satisfying the dominant-energy condition are simply connected. This implies that connected components of event horizons in stationary spacetimes have $\mathbb{R} \times S^2$ topology.

The discussion of the concepts associated with stationary-black hole spacetimes can be concluded by summarizing the properties of the Schwarzschild and Kerr geometries: the extended Kerr spacetime with $m > a$ is a black hole spacetime with the hypersurface $\{r=r_+\}$ forming a nondegenerate, bifurcate Killing horizon generated by the vector field $X + \omega Y$ and surface gravity given by

$$\kappa = \frac{(m^2 - a^2)^{1/2}}{2m[m + (m^2 - a^2)^{1/2}]}$$

In the case $a=0$, where the angular velocity ω vanishes, X is hypersurface orthogonal and becomes the generator of \mathcal{H} . The bifurcation surface in this case is the totally geodesic 2-sphere, along which the four regions in Figure 1 are joined.

Classification of Stationary Solutions (“No-Hair Theorems”)

We confine attention to the “outside region” of black holes, the DOC. (Except for the degenerate case discussed later, the “inside”(black hole) region is not stationary, so that this restriction already follows from the requirement of stationarity.) For reasons of space, we only consider vacuum solutions; there exists a similar theory for electro-vacuum black holes. (There is a somewhat less developed theory for black hole spacetimes in the presence of nonabelian gauge fields.) In connection with a collapse scenario, the vacuum condition begs the question: collapse of what? The answer is twofold: first, there are large classes of solutions of Einstein equations describing pure gravitational waves. It is believed that sufficiently strong such solutions will form black holes. (Whether or not they will do that is related to the cosmic censorship conjecture, *see* Spacetime Topology, Casual Structure and Singularities.) Consider, next, a dynamical situation in which matter is initially present. The conditions imposed in this section correspond then to a final state in which matter has either been radiated away to infinity, or has been swallowed by the black hole (as in the spherically symmetric Oppenheimer–Snyder collapse described above).

Based on the facts below, it is expected that the DOCs of appropriately regular, stationary, vacuum black holes are isometrically diffeomorphic to those of Kerr black holes:

1. *The rigidity theorem* (Hawking). Event horizons in regular, nondegenerate, stationary, analytic vacuum black holes are either Killing horizons for X , or there exists a second Killing vector in $\langle\langle\mathcal{M}\rangle\rangle$.
2. *The Killing horizons theorem* (Sudarsky–Wald). Nondegenerate stationary vacuum black holes such that the event horizon is the union of Killing horizons of X are static.
3. The Schwarzschild black holes exhaust the family of static regular vacuum black holes (Israel, Bunting – Masood-ul-Alam, Chruściel).
4. The Kerr black holes satisfying

$$m^2 > a^2 \quad [11]$$

exhaust the family of nondegenerate, stationary-axisymmetric, vacuum, connected black holes. Here m is the total Arnowitt–Deser–Misner (ADM) mass, while the product am is the total ADM angular momentum. (Of course, these quantities generalize the constants a and m appearing in the Kerr metric.) The framework for the proof has been set up by Carter, and the statement above is due to Robinson.

The above results are collectively known under the name of no-hair theorems, and they have not provided the final answer to the problem so far. There are no *a priori* reasons known for the analyticity hypothesis in the rigidity theorem. Further, degenerate horizons have been completely understood in the static case only.

Yet another key open question is that of the existence of nonconnected regular stationary-axisymmetric vacuum black holes. The following result is due to Weinstein: let $\partial\mathcal{S}_a$, $a=1, \dots, N$, be the connected components of $\partial\mathcal{S}$. Let $X^b = g_{\mu\nu}X^\mu dx^\nu$, where X^μ is the Killing vector field which asymptotically approaches the unit normal to \mathcal{S}_{ext} . Similarly, set $Y^b = g_{\mu\nu}Y^\mu dx^\nu$, Y^μ being the Killing vector field associated with rotations. On each $\partial\mathcal{S}_a$, there exists a constant ω_a such that the vector $X + \omega_a Y$ is tangent to the generators of the Killing horizon intersecting $\partial\mathcal{S}_a$. The constant ω_a is called the angular velocity of the associated Killing horizon. Define

$$m_a = -\frac{1}{8\pi} \int_{\partial\mathcal{S}_a} *dX^b \quad [12]$$

$$L_a = -\frac{1}{4\pi} \int_{\partial\mathcal{S}_a} *dY^b \quad [13]$$

Such integrals are called Komar integrals. One usually thinks of L_a as the angular momentum of each connected component of the black hole. Set

$$\mu_a = m_a - 2\omega_a L_a \quad [14]$$

Weinstein shows that one necessarily has $\mu_a > 0$. The problem at hand can be reduced to a harmonic-map equation, also known as the Ernst equation, involving a singular map from \mathbb{R}^3 with Euclidean metric δ to the two-dimensional hyperbolic space. Let $r_a > 0$, $a = 1, \dots, N-1$, be the distance in \mathbb{R}^3 along the axis between neighboring black holes as measured with respect to the (unphysical) metric δ . Weinstein proved that for nondegenerate regular black holes the inequality [11] holds, and that the metric on $\langle\langle \mathcal{M} \rangle\rangle$ is determined up to isometry by the $3N-1$ parameters

$$(\mu_1, \dots, \mu_N, L_1, \dots, L_N, r_1, \dots, r_{N-1}) \quad [15]$$

just described, with $r_a, \mu_a > 0$. These results by Weinstein contain the no-hair theorem of Carter and Robinson as a special case. Weinstein also shows that, for every $N \geq 2$ and for every set of parameters [15] with $\mu_a, r_a > 0$, there exists a solution of the problem at hand. It is known that for some sets of parameters [15] the solutions will have “strut singularities” between some pairs of neighboring black holes, but the existence of the “struts” for all sets of parameters as above is not known, and is one of the main open problems in our understanding of stationary-axisymmetric electrovacuum black holes. The existence and uniqueness results of Weinstein remain valid when strut singularities are allowed in the metric at the outset, although such solutions do not fall into the category of regular black holes discussed here.

Acknowledgments

This work was partially supported by a Polish Research Committee grant 2 P03B 073 24 and by the Erwin Schroedinger Institute, Vienna.

See also: Asymptotic Structure and Conformal Infinity; Black Hole Mechanics; Critical Phenomena in Gravitational Collapse; Einstein Equations: Exact Solutions; Einstein Equations: Initial Value Formulation; Geometric Flows and the Penrose Inequality; Spacetime Topology, Causal Structure and Singularities.

Further Reading

- Carter B (1973) Black hole equilibrium states. In: de Witt C and de Witt B (eds.) *Black Holes*, Proceedings of the Les Houches Summer School. New York: Gordon and Breach.
- Chruściel PT (2002) Black holes. In: Friedrich H and Frauendiener J (eds.) *Proceedings of the Tübingen Workshop on the Conformal Structure of Space-Times*, Springer Lecture Notes in Physics, vol. 604, pp. 61–102 (gr-qc/0201053). Springer: Heidelberg.
- Hawking SW and Ellis GFR (1973) *The Large Scale Structure of Space-Time*. Cambridge: Cambridge University Press.
- Heusler M (1996) *Black Hole Uniqueness Theorems*. Cambridge: Cambridge University Press.
- Heusler M (1998) Stationary black holes: uniqueness and beyond. *Living Reviews Relativity* 1 (<http://www.livingreviews.org/lrr-1998-6>).
- Nicolas J-P (2002) Dirac fields on asymptotically flat space-times. *Dissertationes Mathematicae* 408: 1–85.
- O’Neill B (1995) *The Geometry of Kerr Black Holes*. Wellesley, MA: A.K. Peters.
- Volkov MS and Gal’tsov DV (1999) Gravitating non-abelian solitons and black holes with Yang–Mills fields. *Physics Reports* 319: 1–83 (hep-th/9810070).
- Wald RM (1984) *General Relativity*. Chicago: University of Chicago Press.

Stationary Phase Approximation

J J Duistermaat, Universiteit Utrecht, Utrecht, The Netherlands

© 2006 Elsevier Ltd. All rights reserved.

Introduction

An oscillatory integral is an integral of the form

$$I(\omega) = \int e^{i\omega\varphi(\theta)} a(\theta) d\theta \quad [1]$$

Here the integration is over a smooth k -dimensional manifold Θ which is provided with a smooth density

$d\theta$. The real variable ω plays the role of a frequency variable, whereas the real-valued smooth function φ on Θ is called the phase function. The amplitude function a is assumed to be a compactly supported complex (vector-) valued smooth function on Θ . The topic of this article is the asymptotic behavior of the oscillatory integral $I(\omega)$ as the frequency ω tends to infinity.

When the manifold Θ is not compact and the amplitude function is not compactly supported, then a smooth cutoff function may be used to write the integral as the sum of an integral with a compactly supported amplitude and one with an amplitude which is equal to zero in a large compact subset of Θ . The

latter integral can be studied if suitable assumptions are made about the asymptotic behavior of the phase function and the amplitude at infinity, but this is not the subject of this article. The use of the exponential function with purely imaginary argument instead of the sine and the cosine is just a matter of convenience.

The first observation about oscillatory integrals in the next section is the principle of stationary phase, which states that the contributions to the integral which are not rapidly decreasing as $\omega \rightarrow \infty$ only come from the stationary points of φ , the points $\theta \in \Theta$ where the total derivative $d\varphi(\theta)$ of φ is equal to zero. This principle is closely related to the observation that a superposition of waves is maximal at points where the waves are in phase, an observation which goes back to Huygens (1690).

Assume that θ_0 is a nondegenerate stationary point of φ . That is, $d\varphi(\theta_0) = 0$ and the Hessian $D^2\varphi(\theta_0)$ of φ at θ_0 is nondegenerate. Then θ_0 is an isolated stationary point of φ , and the contribution to $I(\omega)$ of a neighborhood of θ_0 has an asymptotic expansion of the form

$$I(\omega) \sim e^{i\omega\varphi(\theta_0)} \sum_{r=0}^{\infty} c_r \omega^{-k/2-r}, \quad r \rightarrow \infty$$

Here the leading coefficient c_0 is the product of $a(\theta_0)$ with a nonzero constant which only depends on $D^2\varphi(\theta_0)$ and the density $d\theta$ at θ_0 . For increasing r the coefficients c_r depend on the derivatives of φ and a at θ_0 of increasing order (see the section “The method of stationary phase”).

Usually, even if all the objects are analytic in a neighborhood of θ_0 , the asymptotic power series does not converge. However, there are exceptional cases where the stationary phase approximation is exact. Assume, for instance, that Θ is a compact manifold provided with a symplectic form σ , φ is the Hamiltonian function of a Hamiltonian circle action on Θ with isolated fixed points, and $a(\theta) d\theta = \sigma^k/k!$. Then the stationary points of φ are the fixed points of the circle action, each stationary point of φ is nondegenerate and $I(\omega)$ is equal to the sum over the finitely many stationary points of only the leading terms of the asymptotic expansions at the stationary points. This Duistermaat–Heckman formula is a consequence of a more general localization formula in equivariant cohomology (see the section “Exact stationary phase”).

For the purpose of applications, but also in the analysis of oscillatory integrals, it is worthwhile to allow complex-valued phase functions, but with a local minimum for the imaginary part at the stationary point θ_0 of the real parts. That is, the real part of the exponent $i\omega\varphi(\theta)$ has a local maximum at θ_0 . An extreme case occurs when

$\varphi(\theta) = i\psi(\theta)$ for a real-valued function ψ which has a nondegenerate local minimum at θ_0 , in which case the integrand is a sharply peaking Gaussian density at θ_0 . When φ and a are analytic near θ_0 , then the method of steepest descent consists of deforming the path of integration in the complex domain in such a way that the integrand becomes such a sharply peaking Gaussian density. During the deformation, the integral does not change because of Cauchy’s integral theorem.

An important extension of the theory occurs if the real-valued phase function and the amplitude are allowed to depend smoothly on additional parameters x , which vary in an n -dimensional smooth manifold M . The amplitude is also allowed to depend on ω , with an asymptotic expansion of the form

$$a(x, \theta, \omega) \sim \sum_{r=0}^{\infty} a_r(x, \theta) \omega^{m+(k/2)-r} \quad \text{as } \omega \rightarrow \infty \quad [2]$$

The expansion is supposed to be locally uniformly in (x, θ) and to allow termwise differentiations of any order with respect to the variables (x, θ) . Then the integral

$$I(x, \omega) = \int e^{i\omega\varphi(x, \theta)} a(x, \theta, \omega) d\theta$$

is called an oscillatory integral of order m . Here the function $x \mapsto I(x, \omega)$ is viewed as a continuous superposition of the θ -dependent family of oscillatory functions $x \mapsto e^{i\omega\varphi(x, \theta)} a(x, \theta, \omega)$.

The example which formed the point of departure of Airy (1838) is that $e^{i\omega\varphi(x, \theta)} a(x, \theta, \omega)$ is the wave which arrives at the points x in spacetime which is sent out by a point θ on a reflecting mirror. That is, at x one collects (= integrates over Θ) all the waves sent out by the various points θ of the mirror Θ . The main point of the theory, however, is that in great generality the solutions of linear partial differential equations, such as classical wave equations or quantum mechanical Schrödinger equations, can be represented, as functions of x , as oscillatory integrals. This construction has led to decisive progress in the general theory of linear partial differential equations with smoothly varying coefficients.

According to the principle of stationary phase, the main asymptotic contributions to the integral come from the points θ such that $\partial\varphi(x, \theta)/\partial\theta = 0$. The phase function $\varphi \in C^\infty(M \times \Theta)$ is called nondegenerate if the $(n+k) \times k$ -matrix

$$\frac{\partial^2 \varphi(x, \theta)}{\partial(x, \theta) \partial \theta} \quad \text{has rank } k \quad \text{when} \quad \frac{\partial \varphi(x, \theta)}{\partial \theta} = 0 \quad [3]$$

This is the natural condition to ensure that the set

$$S_\varphi := \left\{ (x, \theta) \in M \times \Theta \mid \frac{\partial \varphi(x, \theta)}{\partial \theta} = 0 \right\}$$

is a smooth n -dimensional submanifold of $M \times \Theta$. The condition [3], moreover, implies that the mapping

$$\iota_\varphi: S_\varphi \ni (x, \theta) \mapsto \left(x, \frac{\partial \varphi(x, \theta)}{\partial x} \right) \in T^*M$$

is a smooth immersion from S_φ into the cotangent bundle T^*M of M . Note that $\xi = \partial \varphi(x, \theta) / \partial x$ is coordinate invariantly defined as a linear form on the tangent space $T_x M$ of M at the point x . That is, $\xi \in (T_x M)^*$ = the dual space of $T_x M$, and $(T_x M)^*$ is the fiber of T^*M over x . In classical mechanics, T^*M is the phase space of the position space M , and a linear form ξ on $T_x M$ is called a momentum vector at the position x . If σ denotes the canonical symplectic form on T^*M , then $\iota_\varphi^* \sigma = 0$. The immersion ι_φ locally embeds S_φ onto a smooth n -dimensional submanifold Λ_φ of M , which is a Lagrangian manifold in T^*M , which by definition means that $\iota_\varphi^* \sigma = 0$.

Oscillatory integrals with very different phase functions and amplitudes can define the same ω -dependent functions on M . The theory of Hörmander (1971, section 3.1) says that the germs of the Lagrangian manifolds Λ_φ and Λ_ψ are the same if and only if φ and ψ define the same class of oscillatory integrals. Moreover, every Lagrangian submanifold Λ of T^*M is locally of the form Λ_φ for some nondegenerate phase function φ . In this way, the mapping $\varphi \mapsto \Lambda_\varphi$ defines a bijection between the set of equivalence classes of germs of nondegenerate phase functions and the set of germs of Lagrangian submanifolds of T^*M . Let Λ be an immersed Lagrangian submanifold of T^*M . A global oscillatory integral of order m on M , defined by Λ , is a locally finite sum $u(x, \omega)$ of oscillatory integrals of order m with nondegenerate phase functions φ such that $\Lambda_\varphi \subset \Lambda$. The leading terms of the amplitudes correspond to a section s of a canonically defined complex line bundle λ over Λ , which is called the principal symbol of u (see the section "The principal symbol on the Lagrangian manifold").

If P is a linear partial differential operator, such as the wave operators, in which the coefficients may depend in a smooth way on x and in a polynomial way on ω , then the condition that Pu is asymptotically small implies that $p=0$ on Λ , in which p is a smooth function on T^*M , called the principal symbol of P . Because Λ is a Lagrangian manifold,

the equation $p=0$ implies that Λ is invariant under the flow of the Hamiltonian system with Hamilton function equal to p . Furthermore, the principal symbol s of u satisfies a homogeneous first-order ordinary differential equation along the solution curves of the Hamiltonian system. Conversely, these properties can be used to construct global oscillatory integrals u which asymptotically satisfy $Pu=0$ and have prescribed initial values. This theory, due to Maslov (1972), may be viewed as a far reaching generalization of the WKB method.

Let $\pi: T^*M \rightarrow M: (x, \xi) \mapsto x$ denote the canonical projection from T^*M onto M . The projections into M of the solution curves in a Lagrangian submanifold Λ of T^*M , of a Hamiltonian system which leaves Λ invariant, are the ray bundles of geometrical optics. If Λ is not transversal to the fiber of T^*M at (x, ξ) , then the ray bundle exhibits a caustic at the point $x \in M$, and the oscillatory integral is asymptotically of larger order than ω^m near x . Applying the theory of unfoldings of singularities to the phase function, one can determine the structurally stable caustics and obtain normal forms of the oscillatory integrals in the structurally stable cases (see the section "Caustics").

If we also integrate over the frequency variable ω , then we obtain the Fourier integral distributions u of Hörmander (1971, sections 1.2 and 3.2). In this case the corresponding Lagrangian manifold is conic in the sense that if $(x, \xi) \in \Lambda$, then $(x, \tau \xi) \in \Lambda$ for every $\tau > 0$. The wave front set of u , which is the microlocal singular locus of the distribution u , is contained in Λ , with equality if the principal symbol of u is not equal to zero at the corresponding stationary points of the phase function. Fourier integral operators are defined as the linear operators acting on distributions, of which the distribution kernels are Fourier integral distributions. Under a suitable transversality condition for the Lagrangian manifolds of the distribution kernels, the composition of two Fourier integral operators is again a Fourier integral operator, and the principal symbol of the composition is a product of the principal symbols. The proof is an application of the method of stationary phase. Fourier integral operators are a very powerful tool in the analysis of linear partial differential operators with smoothly varying coefficients (see Hörmander (1985)).

The Principle of Stationary Phase

The principle of stationary phase says that if the phase function φ has no stationary points in the support of the amplitude function a , then the

oscillatory integral [1] is rapidly decreasing, in the sense that for every N we have $I(\omega) = O(\omega^{-N})$ as $\omega \rightarrow \infty$. For the proof, one introduces a vector field v on Θ such that $v\varphi = 1$ on a neighborhood of the support of a . Then $e^{i\omega\varphi} = (i\omega)^{-1}v(e^{i\omega\varphi})$, and an integration by parts in [1] yields that

$$I(\omega) = \frac{1}{i\omega} \int e^{i\omega\varphi(\theta)} ({}^tva)(\theta) d\theta$$

where tv denotes the transposed of the linear partial differential operator v . Iterating this, the rapid decrease of $I(\omega)$ follows.

Using cutoff functions, $I(\omega)$ is, modulo a rapidly decreasing function, equal to an oscillatory integral with phase function φ and an amplitude which has support in an arbitrarily small neighborhood of the set of stationary points of φ . In this sense, the contributions to the integral which are not rapidly decreasing come only from the stationary points of φ .

The Method of Stationary Phase

Assume that θ_0 is a nondegenerate stationary point of φ . Then θ_0 is an isolated stationary point of φ . Using local coordinates near θ_0 , the contribution to [1] from the neighborhood of θ_0 can be written as an oscillatory integral with $\Theta = \mathbb{R}^k$ and a phase function φ which has a nondegenerate stationary point at 0. Write $Q = D^2\varphi(0)$. According to the Morse lemma, there is smooth substitution of variables $\theta = T(y)$ such that $T(0) = 0$, $DT(0) = I$, and $\varphi(T(y)) = \varphi(0) + \langle Qy, y \rangle / 2$ for all y in a neighborhood of 0 in \mathbb{R}^k . Applying this substitution of variables to [1] we obtain

$$I(\omega) = e^{i\omega\varphi(0)} \int_{\mathbb{R}^k} e^{i\omega\langle Qy, y \rangle / 2} b(y) dy$$

where b is a compactly supported smooth function on \mathbb{R}^k with $b(0) = a(0)$. Now the Fourier transform of the function $y \mapsto e^{i\omega\langle Qy, y \rangle / 2}$ is equal to the function

$$\eta \mapsto \left(\det \left(\frac{\omega}{2\pi i} Q \right) \right)^{-1/2} e^{i\omega^{-1} \langle Q^{-1}\eta, \eta \rangle / 2} \quad [4]$$

Both in the definition of the square root of the determinant and in the proof one uses the analytic continuation to the domain of complex-valued symmetric bilinear forms Q for which the imaginary part of Q is positive definite. For purely imaginary Q we have the familiar formula for the Fourier transform of a Gaussian density (see Hörmander

(1990, theorem 7.6.1)). The Taylor expansion of the exponential factor in [4] then yields that

$$\begin{aligned} & \int_{\mathbb{R}^k} e^{i\omega\langle Qy, y \rangle / 2} b(y) dy \\ & \sim \left(\det \left(\frac{\omega}{2\pi i} Q \right) \right)^{-1/2} \sum_{r=0}^{\infty} (-2i\omega)^{-r} \frac{1}{r!} \\ & \quad \times \left\langle Q^{-1} \frac{\partial}{\partial y}, \frac{\partial}{\partial y} \right\rangle^r b(y) \Big|_{y=0} \end{aligned}$$

as $\omega \rightarrow \infty$ (see Hörmander (1990, lemma 7.7.3)).

It is important for the applications that, if the phase function and amplitude depend smoothly on parameters, all the constructions can be made to depend smoothly on the parameters.

Exact Stationary Phase

Suppose that we have given an action of a Lie group G on the manifold Θ . Let \mathfrak{g} denote the Lie algebra of G . For any $g \in G$ and $X \in \mathfrak{g}$ the corresponding diffeomorphism of Θ and vector field on Θ is denoted by g_Θ and X_Θ , respectively. If $\Omega(\Theta)$ denotes the algebra of smooth differential forms on Θ , then we consider the algebra $S\mathfrak{g}^* \otimes \Omega(\Theta)$ of all $\Omega(\Theta)$ -valued polynomials on \mathfrak{g} , where $S\mathfrak{g}^*$ denotes the algebra of all polynomial functions on \mathfrak{g} . On $S\mathfrak{g}^* \otimes \Omega(\Theta)$ we have the action of $g \in G$ which sends α to $X \mapsto g_\Theta^*(\alpha(\text{Ad } g X))$. Let $A = (S\mathfrak{g}^* \otimes \Omega(\Theta))^G$ denote the subalgebra of all G -invariant elements of $S\mathfrak{g}^* \otimes \Omega(\Theta)$. The equivariant exterior derivative D is defined by

$$(D\alpha)(X) = d(\alpha(X)) - i_{X_\Theta}(\alpha(X))$$

If α is homogeneous as a differential form of degree p and homogeneous as a polynomial on \mathfrak{g} of degree q , then $r = p + 2q$ is called the total degree of α . Let A^r denote the space of sums of such $\alpha \in A$ of total degree r . Then $D_r = D: A^r \rightarrow A^{r+1}$ and $D_r \circ D_{r-1} = 0$. The space $H_G^r(\Theta) := \ker D_r / \text{Im } D_{r-1}$ is called the equivariant cohomology in degree r , in the model of Cartan (1950).

Assume that Θ is compact and oriented, and that the action of G preserves the orientation. If $\alpha \in A$, then we denote by $\alpha(X)^{[k]}$ the volume part of the differential form $\alpha(X)$, and

$$\left(\int \alpha \right)(X) := \int_{\Theta} \alpha(X)^{[k]}, \quad X \in \mathfrak{g}$$

defines an $\text{Ad } G$ -invariant function $\int \alpha$ on \mathfrak{g} . Now $\alpha = D\beta$ implies that $\alpha(X)^{[k]}$ is equal to the exterior derivative of $\beta(X)^{[k-1]}$, and therefore $\int \alpha = 0$, in view of Stokes' theorem. It follows that integration over Θ yields a linear mapping \int from $H_G(\Theta)$ to $(S\mathfrak{g}^*)^{\text{Ad } G}$, which is called integration in equivariant cohomology.

Now assume that also the Lie group G is compact, and let $X \in \mathfrak{g}$. Then the zero-set Z_X of X_Θ in Θ has finitely many connected components F , each of which is a smooth and compact submanifold of Θ . In general, the F 's can have different dimensions. The linearization LX of the vector field X_Θ along F acts linearly on the normal bundle NF of F . If Ω is the curvature form of NF , then

$$\varepsilon(X) := \det_{\mathbb{C}} \left[\frac{i}{2\pi} (LX - \Omega) \right]$$

is called the equivariant Euler form of NF . $\varepsilon(X)$ is an invertible element in the algebra $\Omega^{\text{even}}(F)$. The localization formula of Berline–Vergne (1982) and Atiyah–Bott (1984) now says that if $D\alpha = 0$ then

$$\left(\int \alpha \right)(X) = \sum_F \int_F (i_F^* \alpha(X) / \varepsilon(X))^{\dim F}$$

Assume that σ is a symplectic form on Θ , which implies that $k = 2l$ is even. Furthermore, assume that the infinitesimal action of \mathfrak{g} on Θ is Hamiltonian, which means that there exists a G -equivariant smooth mapping $\mu: \Theta \rightarrow \mathfrak{g}^*$, called the momentum mapping, such that $i_{X_\Theta} \sigma = -d(\mu(X))$ for every $X \in \mathfrak{g}$. Here μ is viewed as an element of $(\mathfrak{g}^* \otimes \Omega^0(\Theta))^G \subset A$. Then $\hat{\sigma}(X) := \sigma - \mu(X)$ defines an element $\hat{\sigma} \in A$ such that $D\hat{\sigma} = 0$. In turn, this implies that the form

$$\beta(X) := e^{-i\omega\hat{\sigma}(X)} = e^{i\omega\mu(X)} \sum_{r=0}^l (-i\omega\sigma)^r / r!$$

is equivariantly closed, and the localization formula of equivariant cohomology applied to this case yields the Duistermaat–Heckman (1982, 1983) formula. Because $\beta(X)^{[k]} = e^{i\omega\mu(X)} (-i\omega\sigma)^l / l!$, its integral over Θ is an oscillatory integral with phase function $\mu(X)$. The stationary points of $\mu(X)$ are the zeros of X_Θ and the stationary points of $\mu(X)$ are nondegenerate if and only if the zeros of X_Θ are isolated. It follows that in this case the oscillatory integral is equal to the leading term in the stationary-phase approximation.

The Principal Symbol on the Lagrangian Manifold

Let $u(x, \omega)$ be a global oscillatory integral of order m defined by Λ , and let $(x_0, \xi_0) \in \Lambda$. One way to define the principal symbol of u at $(x_0, \xi_0) \in \Lambda$ is to test u with an oscillatory function of the form $e^{-i\omega\psi(x)} b(x)$, in which $d\psi(x_0) = \xi_0$, the support of b is contained in a small neighborhood of x_0 , and $b(x_0) = 1$. If u is locally represented by the phase function φ and

amplitude a , and $(x_0, \xi_0) = \iota_\varphi(x_0, \theta_0)$, then the phase function $\varphi(x, \theta) - \psi(x)$ in the oscillatory integral

$$\langle u, e^{-i\psi} b \rangle = \int_M \int_\Theta e^{i(\varphi(x, \theta) - \psi(x))} a(x, \theta) b(x) d\theta dx$$

has a stationary point at (x_0, θ_0) , which means that Λ and $d\psi$ intersect at (x_0, ξ_0) . Here the 1-form $d\psi$ on M , which is a section of $\pi: T^*M \rightarrow M$, is viewed as a submanifold of T^*M . Locally the Lagrangian submanifolds of T^*M which are transversal to the fibers of $\pi: T^*M \rightarrow M$ are precisely the manifolds of the form $d\psi$. The stationary point of $\varphi - \psi$ is nondegenerate if and only if $L := T_{(x_0, \xi_0)} \Lambda$ and $L_{\psi} := T_{(x_0, \xi_0)}(d\psi)$ are transversal. In this case, the method of stationary phase can be applied in order to obtain an asymptotic expansion in terms of powers of ω . The coefficient of the leading term of order ω^m depends only on the Lagrangian plane L_ψ , which is transversal to both L and the tangent space of the fiber of T^*M , and not on the other data of ψ and b . If \mathcal{L} denotes the set of all Lagrangian planes in $T_{(x_0, \xi_0)}(T^*M)$ which are transversal to both L and the fiber, then the complex-valued functions on \mathcal{L} which arise in this way form a one-dimensional complex vector space $L_{(x_0, \xi_0)}$. The $L_{(x_0, \xi_0)}$ for $(x_0, \xi_0) \in \Lambda$ form a complex line bundle λ over Λ which is canonically isomorphic to the tensor product of the line bundle of half-densities and the Maslov line bundle, a line bundle with structure group $\mathbb{Z}/4\mathbb{Z}$ (see Duistermaat (1974, section 1.2)). In this way, the principal symbol s of u can be viewed as a section of the line bundle λ over Λ .

Caustics

Let (x_0, ξ_0) be a point in the Lagrangian submanifold Λ of T^*M . The restriction to Λ of the projection $\pi: T^*M \rightarrow M$ is a diffeomorphism from an open neighborhood of (x_0, ξ_0) in Λ onto an open neighborhood of x_0 in M , if and only if Λ is transversal to the fiber of T^*M at (x_0, ξ_0) . If $\Lambda = \Lambda_\varphi$ for a nondegenerate phase function φ , $(x_0, \xi_0) \in S_\varphi$ and $(x_0, \xi_0) = \iota_\varphi(x_0, \theta_0)$, then this condition is in turn equivalent to the condition that θ_0 is a nondegenerate stationary point of $\theta \mapsto \varphi(x_0, \theta)$. An application of the method of stationary phase shows that in this case the oscillatory integral is equal to a progressing wave of the form $e^{i\omega\psi(x)} b(x, \omega)$. Here $\psi(x) = \varphi(x, \theta(x))$, where $\theta(x)$ is the stationary point of $\theta \mapsto \varphi(x, \theta)$, and $b(x, \omega)$ has an asymptotic expansion as in [2] with $k = 0$.

If θ_0 is a degenerate stationary point of $\theta \mapsto \varphi(x_0, \theta)$ and $a_0(x_0, \theta_0) \neq 0$, then the oscillatory integral is not of order $O(\omega^m)$. That is, it is of larger order than at points where we have a nondegenerate

stationary point. For this reason, the points (x_0, ξ_0) at which Λ is not transversal to the fibers of $\pi: T^*M \rightarrow M$ are called the caustic points of Λ . Their projections $x_0 \in M$ form the caustic set in M .

In the theory of unfoldings of singularities, the germs of the families of functions $x \mapsto (\theta \mapsto \varphi(x, \theta))$ and $y \mapsto (\mu \mapsto \psi(y, \mu))$ are called equivalent if there exists a germ of a diffeomorphism of the form $H: (x, \theta) \mapsto (y(x), \mu(x, \theta))$ and a smooth function $\chi(x)$ such that $\psi(y(x), \mu(x, \theta)) = \varphi(x, \theta) + \chi(x)$. If $J(y, \omega)$ is an oscillatory integral with phase function ψ , integration variable μ and parameter y , then the substitution of variables $\mu = \mu(x, \theta)$ in the integral, followed by the substitution of variables $y = y(x)$ in the parameters, yields that $J(y, \omega) = e^{i\omega\chi(x)} I(x, \omega)$, in which $I(x, \omega)$ is an oscillatory integral with phase function φ and an amplitude function of the same order as the amplitude function of J . The germ φ is called stable if every nearby germ ψ is equivalent to φ . The Morse lemma with parameters implies that this is the case if $\theta \mapsto \varphi(x_0, \theta)$ has a nondegenerate stationary point at θ_0 . However, the theory of unfoldings of singularities of Thom and Mather shows that there are many stable germs with degenerate critical points. Moreover, in dimension $n \leq 5$ the generic germ is stable, and is equivalent to a germ in a finite list of normal forms.

The simplest example of a normal form with degenerate critical points is $\varphi(x, \theta) = \theta^3 + x_1\theta$. Here we have taken $k=1$, but still allowed an arbitrary dimension $n \geq 1$ of M . In this normal form, the stationary points correspond to $3\theta^2 + x_1 = 0$, which is a manifold which over the x -space folds over at $x_1 = 0$. The stationary point is degenerate if and only if $6\theta = 0$, hence $x_1 = 0$, which means that $x_1 = 0$ is the caustic set. If the amplitude is equal to 1, then the oscillatory integral is equal to $\omega^{-1/3} \text{Ai}(\omega^{2/3}x_1)$, in which $\text{Ai}(z)$ denotes the Airy function. If the amplitude is nonzero at a degenerate critical point, then the oscillatory integral near the corresponding caustic point is asymptotically of the same order as $\omega^{-1/3} \text{Ai}(\omega^{2/3}x_1)$, which implies that the oscillatory integral is a factor $\omega^{1/6}$ larger at these caustic points than at the points away from the caustic set. In Airy (1838), where the Airy function was introduced, Airy considered light in a neighborhood of a caustic as an oscillatory integral. Then, under suitable genericity conditions, he brought the phase function into the normal form $\theta^3 + x_1\theta$. Even for stable normal forms in low dimensions, the interference patterns near the

caustic points can be very intricate (see, e.g., Berry *et al.* (1979)). A survey of the application of the theory of unfoldings to caustics in oscillatory integrals can be found in Duistermaat (1974).

See also: Equivariant Cohomology and the Cartan Model; Feynman Path Integrals; Functional Integration in Quantum Physics; Hamiltonian Group Actions; h -Pseudodifferential Operators and Applications; Multiscale Approaches; Normal Forms and Semiclassical Approximation; Optical Caustics; Path Integrals in Noncommutative Geometry; Perturbation Theory and its Techniques; Schrödinger Operators; Singularity and Bifurcation Theory; Wave Equations and Diffraction.

Further Reading

- Airy GB (1838) On the intensity of light in a neighborhood of a caustic. *Transactions of the Cambridge Philosophical Society* 6: 379–403.
- Atiyah MF and Bott R (1984) The moment map and equivariant cohomology. *Topology* 23: 1–28.
- Berline N and Vergne M (1982) Classes caractéristiques équivariantes. Formules de localisation en cohomologie équivariante. *Comptes Rendus Hebdomadaires des Seances de l'Academie des Sciences, Paris* 295: 539–541.
- Berry MV, Nye JF, and Wright FJ (1979) The elliptic umbilic diffraction catastrophe. *Philosophical Transactions of the Royal Society of London* A291: 453–484.
- Cartan H (1950) Notion d'algèbre différentielle; applications ou opère un groupe de Lie, and: La transgression dans un groupe de Lie et dans un fibré principal. In: *Colloque de Topologie*, pp. 15–27, 57–71 Bruxelles: C.B.R.M.
- Duistermaat JJ (1974) Oscillatory integrals, Lagrange immersions and unfoldings of singularities. *Communications in Pure and Applied Mathematics* 27: 207–281.
- Duistermaat JJ and Heckman GJ (1982) On the variation in the cohomology of the symplectic form of the reduced phase space. *Inventiones Mathematicae* 69: 259–268.
- Duistermaat JJ and Heckman GJ (1983) On the variation in the cohomology of the symplectic form of the reduced phase space. *Inventiones Mathematicae* 72: 153–158.
- Hörmander L (1971) Fourier integral operators I. *Acta Mathematica* 127: 79–183.
- Hörmander L (1983, 1990) *The Analysis of Linear Partial Differential Operators I*. Berlin: Springer.
- Hörmander L (1985) *The Analysis of Linear Partial Differential Operators IV*. Berlin: Springer.
- Huygens C (1690) *Traité de la Lumière*, Leyden: Van der Aa. (English translation: *Treatise on Light*, 1690, reprint by Dover Publications, New York, 1962.)
- Maslov VP (1965) *Perturbation Theory and Asymptotic Methods* (In Russian), Moscow: Moskov. Gos. University. (French translation by Lascoux J and Sénéor R (1972) Paris: Dunod.)

Statistical Mechanics and Combinatorial Problems

R Zecchina, International Centre for Theoretical Physics (ICTP), Trieste, Italy

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Equilibrium statistical mechanics and combinatorial optimization – which is viewed here as a branch of discrete mathematics and theoretical computer science – have common roots. Phase transition are mathematical phenomena which are not limited to physical systems but are typical of many combinatorial problems, one famous example being the percolation transition in random graphs. Similarly, the understanding of relevant physical problems, such as three-dimensional lattice statistics or two-dimensional quantum statistical mechanics problems, is strictly related to the question of purely combinatorial origin of solving counting problems over nonplanar lattices. Most of the tools and concepts which have allowed to solve problems in one field have a natural counterpart in the other. While the possibility of solving exactly physical models is always connected to the presence of some algebraic properties which guarantee integrability, in the combinatorial approach the emphasis is more on algorithms that can be applied to problem instances in which the symmetries behind integrability might be absent. Also at the level of out-of-equilibrium phenomena, there exists a deep connection between physics and combinatorics: just like physical processes, local algorithms have to deal with an exponentially large set of possible configurations and their out-of-equilibrium analysis constitutes a theory of how problems are actually solved.

Computational complexity theory deals with classifying problems in terms of the computational resources, typically time, required for their solution. What can be measured (or computed) is the time that a particular algorithm uses to solve the problem. This time in turn depends on the implementation of the algorithm as well as on the computer the program is running on. The theory of computational complexity provides us with a notion of complexity that is largely independent of implementational details and the computer at hand. This is not surprising, since it is related to a highly nontrivial question, that is: what do we mean by saying that a combinatorial problem is solvable?

Problems which can be solved in polynomial time are considered to be tractable and compose the so-called polynomial (P) class. The harder problems are

grouped in a larger class called NP, where NP stands for “nondeterministic polynomial time.” These problems are such that a potential solution can be checked rapidly in polynomial time, while finding a solution may require exponential time in the worst case. In turn, the hardest problems in NP belong to a subclass called NP-complete: an efficient algorithm for solving one NP-complete problem could be easily modified to effectively solve any problem in NP. By now, a huge number of NP-complete problems has been identified, and the lack of such an algorithm corroborates the widespread conjecture $P \neq NP$, that is, that no such algorithm exists. However, NP-complete problems are not always hard: when their resolution complexity is measured with respect to some underlying probability distribution of problem instances, NP-complete problems are often easy to solve on average. To deepen the understanding of the average-case complexity (and of the huge variability of running times observed in numerical experiments), computer scientists, mathematicians, and physicists have focused their attention on the study of random instances of hard combinatorial problems, seeking for a link between the onset of exponential-time complexity and some intrinsic (i.e., algorithm independent) properties of the randomized NP-complete problems. These types of questions have merged combinatorial optimization with statistical physics of disordered systems.

Computational complexity theory can also be formulated for counting problems: similarly to optimization problems, equivalence classes can be defined which separate polynomially solvable counting problems with the hard ones – the so-called #P and #P-complete problems. Complexity theory for counting problems makes the connections with statistical mechanics even more direct in that counting solutions is nothing but a computation of a partition function.

Two simple theorems by Jerrum and Sinclair (1989) (which can be easily extended to many combinatorial problems) can help in clarifying these connections.

The first theorem tells us that any randomized algorithm (e.g., Monte Carlo) for approximating the partition function of a generic spin glass model – the so-called spin glass problem – could be used to solve all the other NP combinatorial problems. The second theorem tells us that an algorithm for evaluating exactly the partition function of the ferromagnetic Ising model over a general graph would again solve any other problem in the class #P, which, as mentioned above, is the generalization of

the NP class to counting problems and obviously contains the class NP as a particular case.

Let us consider the following slightly simplified definition of the Ising and the spin glass problems.

Problem instance A symmetric matrix J_{ij} with entries in $\{-1, 0, 1\}$ and an inverse temperature β .

Output The partition function $Z = \sum_{\{\sigma_i\}} 2^{-\beta H(\sigma)}$, where $H(\sigma) = -\sum_{i,j} J_{ij} \sigma_i \sigma_j$ with $\sigma_i = \pm 1$.

Moreover, let us define the fully polynomial randomized approximation scheme (FPRAS) for counting and decision problems. A FPRAS for a function f from problem instances to real numbers is a probabilistic algorithm that in polynomial time in the problem size n and in the relative error $\epsilon \in [0, 1]$, outputs with high probability a number which approximates $f(n)$ within a ratio $1 + \epsilon$. Given the above definitions, the theorems can be stated as follows:

Theorem 1 *There can be no FPRAS for the spin glass problem unless $P=NP$, that is, all problems in NP turn out to be solvable in polynomial time.*

Theorem 2 *The Ising problem is #P-complete even when the matrix J_{ij} is non-negative, that is, an algorithm which outputs in polynomial time the exact Ising partition function for an arbitrary graph could be used to solve any other counting problem in #P.*

The above theorems hold for arbitrary graphs, in particular for those graph or lattice realizations which are particularly hard to analyze, the so-called worst cases. There exist no similar proofs of computational hardness for more restricted and realistic structures, such as, for instance, three-dimensional regular lattices for the Ising problem or finite connectivity random graphs for spin glasses.

As a final introductory remark, it is worth mentioning that the connections between worst-case complexity and the average case one is the building block of modern cryptography and communication theory. On the one hand, the so-called RSA cryptosystem is based on factoring large integers, a problem which is believed to be hard on average while it is not known to be so in the worst case. On the other hand, alternative cryptographic systems have been proposed which rely on a worst-case/average-case equivalence (see, e.g., the theorem of Ajtai (1996) concerning some hidden vector problems in high-dimensional lattices.)

As far as communication theory is concerned, average-case complexity is indeed crucial: while Shannon's theorem (1948) provides a very general result stating that many optimal codes do exist (in

fact, random codes are optimal), the decoding problem is in general NP-complete and therefore potentially intractable. However, since the choice of the coding scheme is part of the design, what matters are the average-case behavior of the decoding algorithm (and its large deviations) and very efficient codes which can solve on average the decoding problem close to Shannon's bounds are known.

In what follows, we will limit the discussion to two basic examples of combinatorial and counting problems which are representative and central to both computer science and statistical physics.

Constraint Satisfaction Problems

Combinatorial problems are usually written as constraint satisfaction problems (CSPs): n discrete variables are given which have to satisfy m constraints, all at the same time. Each constraint can take different forms depending on the problem under study: famous examples are the K -satisfiability (K -SAT) problem in which constraints are an "OR" function of K variables in the ensemble (or their negations) and the graph Q -coloring problem in which constraints simply enforce the condition that the endpoints of the edges in the graph must not have the same color (among the Q possible ones). Quite in general a generic CSP can be written as the problem of finding a zero-energy ground state of an appropriate energy function and its analysis amounts at performing a zero-temperature statistical physics study. Hard combinatorial problems are those which correspond to frustrated physical model systems.

Given an instance of a CSP, one wants to know whether there exists a solution, that is, an assignment of the variables which satisfies all the constraints (e.g., a proper coloring). When it exists, the instance is called SAT, and one wants to find a solution. Most of the interesting CSPs are NP-complete: in the worst case, the number of operations needed to decide whether an instance is SAT or not is expected to grow exponentially with the number of variables. But recent years have seen an upsurge of interest in the theory of typical-case complexity, where one tries to identify random ensembles of CSPs which are hard to solve, and the reason for this difficulty. As already mentioned, random ensembles of CSPs are also of great theoretical and practical importance in communication theory, since some of the best modern error-correcting codes (the so-called low-density parity check codes) are based on such constructions.

Satisfiability and Spin Glass Models

The archetypical example of CSP is satisfiability (SAT). This is a core problem in computational complexity: it is the first one to have been shown NP-complete, and since then thousands of problems have been shown to be computationally equivalent to it. Yet it is not so easy to find difficult instances. The main ensemble which has been used for this goal is the random K -SAT ensemble (for $K > 2$, K -SAT is NP-complete).

The SAT problem is defined as follows. Given a vector of $\{0, 1\}$ Boolean variables $\mathbf{x} = \{x_i\}_{i \in I}$, where $I = \{1, \dots, n\}$, consider a SAT formula defined by

$$\mathcal{F}(\mathbf{x}) = \bigwedge_{a \in A} C_a(\mathbf{x})$$

where A is an arbitrary finite set (disjoint with I) labeling the clauses C_a ; $C_a(\mathbf{x}) = \bigvee_{i \in I(a)} J_{a,i}(x_i)$; any literal $J_{a,i}(x_i)$ is either x_i or $\sim x_i$ (“not” x_i); and finally, $I(a) \subset I$ for every $a \in A$. Similarly to $I(a)$, we can define the set $A(i) \subset A$ as $A(i) = \{a : i \in I(a)\}$, that is, the set of clauses containing variable x_i or its negation.

Given a formula \mathcal{F} , the problem of finding a variable assignment \mathbf{s} such that $\mathcal{F}(\mathbf{s}) = 1$, if it exists, can also be written as a spin glass problem as follows: if we consider a set of n Ising spins, $\sigma_i \in \{\pm 1\}$ in place of the Boolean variables ($\sigma_i = -1, 1 \leftrightarrow x_i = 0, 1$) we may write the energy function associated to each clause as follows:

$$E_a = \prod_{r=1}^K \frac{(1 + J_{a,i_r} \sigma_{i_r})}{2}$$

where $J_{a,i} = -1$ (resp. $J_{a,i} = 1$) if x_i (resp. \tilde{x}_i) appears in clause a . The total energy of a configuration $E = \sum_{a=1}^{|A|} E_a$ is nothing but a K -spin spin glass model.

Random K -SAT is a version of SAT in which each clause is taken to involve exactly K distinct variables, randomly chosen and negated with uniform distribution. Its energy function corresponds to a spin glass system over a finite connectivity (diluted) random graph.

In recent years random K -SAT has attracted much interest in computer science and in statistical physics. The interesting limit is the thermodynamic limit $n \rightarrow \infty$, $m = |A| \rightarrow \infty$ at fixed clause density $\alpha = m/n$.

Its most striking feature is certainly its sharp threshold. It is strongly believed that there exists a phase transition for this problem: numerical and

heuristic analytical arguments are in support of the so-called satisfiability threshold conjecture:

Conjecture There exists $\alpha_c(K)$ such that with high probability:

- if $\alpha < \alpha_c(K)$, a random instance is satisfiable;
- if $\alpha > \alpha_c(K)$, a random instance is unsatisfiable.

Although this conjecture remains unproven, the existence of a nonuniform sharp threshold has been established by Friedgut (1997). A lot of effort has been devoted to understanding this phase transition. This is interesting both from physics and the computer science points of view, because the random instances with α close to α_c are the hardest to solve. There exist rigorous results that give bounds for the threshold $\alpha_c(K)$: using these bounds, it was shown that $\alpha_c(K)$ scales as $2^K \ln(2)$ when $K \rightarrow \infty$.

On the statistical physics side, the cavity method (which is the generalization to disordered systems characterized by ergodicity breaking of the iterative method used to solve exactly physical models on the Bethe lattice), is a powerful tool which is claimed to be able to compute the exact value of the threshold, giving for instance $\alpha_c(3) \simeq 4.2667 \dots$. It is a non-rigorous method but the self-consistency of its results have been checked by a “stability analysis,” and it has also led to the development of a new family of algorithms – the so-called “survey propagation” – which can solve efficiently very large instances at clause densities which are very close to the threshold (for technical details see Mézard and Zecchina (2002) and Braunstein *et al.* (2005) and references therein).

The main hypothesis on which the cavity analysis of random K -SAT relies is the existence, in a region of clause density $[\alpha_d, \alpha_c]$ close to the threshold, of an intermediate phase called the “hard-SAT” phase; see Figure 1. In this phase the set \mathcal{S} of solutions (a subset of the vertices in an n -dimensional hypercube) is supposed to split into many disconnected clusters $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \dots$. If one considers two solutions X, Y in the same cluster \mathcal{S}_j , it is

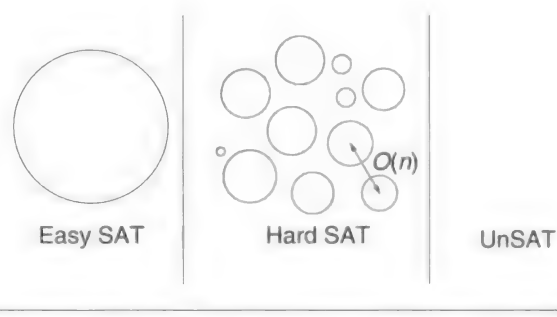


Figure 1 A pictorial representation of the clustering transition in random K -SAT.

possible to walk from X to Y (staying in S) by flipping at each step a finite number of variables. If, on the other hand, X and Y are in different clusters, in order to walk from X to Y (staying in S), at least one step will involve an extensive number (i.e., $\propto n$) of flips. This clustered phase is held responsible for entrapping many local-search algorithms into non-optimal metastable states. This phenomenon is not exclusive to random K -SAT. It is also predicted to appear in many other hard SAT and optimization problems such as “coloring,” and corresponds to the so called “one-step replica symmetry breaking” (1RSB) phase in the language of statistical physics. It is also a crucial limiting feature for decoding algorithms in some error correcting codes.

The only CSP for which the existence of the clustering phase has been established rigorously is the polynomial problem of solving random linear equation in GF (Motwani and Raghavan 2000). For random K -SAT, rigorous probabilistic bounds can be used to prove the existence of the clustering phenomenon, for large enough K , in some region of α included in the interval $[\alpha_d(K), \alpha_c(K)]$ predicted by the statistical physics analysis.

In the analysis of CSP like K -SAT, two main questions are in order. The first is of algorithmic nature and asks for an algorithm which decides whether for a given CSP instance all the constraints can be simultaneously satisfied or not. The second question is more theoretical and deals with large random instances, for which one wants to know the structure of the solution space and predict the typical behavior of classes of algorithms.

Message-Passing Algorithms from Statistical Physics

The algorithmic contributions of statistical mechanics to combinatorial optimization are numerous and important (a representative example being the celebrated “simulated annealing algorithm”). For the sake of brevity, here we limit the discussion to the so-called “message-passing algorithms” which are also of great interest in coding theory.

The statistical analysis of the cavity equations allows to study the average properties of ensemble of problems and it is totally equivalent to the replica method in which the average over the ensemble is the first step in any calculation. The survey propagation (SP) equations are a formulation of the cavity equations which is valid for each specific instance and is able to provide information about the statistical behavior of the individual variables in the stable and metastable states of a given energy density (i.e., given fraction of violated constraints).

The single-sample SP equations are nicely described in terms of the factor graph representation used in information theory to characterize error-correcting codes. In the factor graph, the N variables i, j, k, \dots are represented by circular “variable nodes,” whereas the M clauses a, b, c, \dots are represented by square “function nodes.” For random K -SAT, the function nodes have connectivity K , while the variable nodes have an average Poisson connectivity $K\alpha$.

The iterative SP equations are examples of message-passing procedures. In message-passing algorithms such as the so-called “belief propagation (BP) algorithm” used in error-correcting codes and statistical inference problems, the unknowns which are self-consistently evaluated by iteration are the marginals over the solution space of the variables characterizing the combinatorial problem (the probability space is the set of all solutions sampled with uniform measure). According to the physical interpretation, the quantities that are evaluated by SP are the probability distributions of local fields over the set of clusters. That is, while BP performs a “white” average over solutions, SP takes care of cluster-to-cluster fluctuations, telling us which is the probability of picking up a cluster at random and finding a given variable biased in a certain direction (or unfrozen if it is paramagnetic in the cluster). SP computes quantities which are probabilities over different pure states: the order parameter which is evaluated as fixed point of the SP equations is a probability measure in a space of functions, or for finite n , the full list of probability densities describing the cluster-to-cluster fluctuations of the variables.

In both SP and BP one assumes knowledge of the marginals of all variables in the temporary absence of one of them and then writes the marginal probability induced on this “cavity” variable in absence of another third variable interacting with it (i.e., the so-called Bethe lattice approximation for the problem). These relations define a closed set of equations for such cavity marginals that can be solved iteratively (this fact is known as message-passing technique). The equations become exact if the cavity variables acting as inputs are uncorrelated. They are conjectured to be an asymptotically exact approximation over random locally tree-like structures such as, for instance, the random K -SAT factor graph. Both BP and SP can be derived in a variational framework.

Complexity of Counting Problems

In order to describe the nature of computational complexity of counting in physical models, it is enough to consider the classical Ising problem. The

computation of the Ising partition function or, more in general, of the weighted matching polynomial, is the root problem of lattice statistics.

For planar graphs like, for example, two-dimensional regular lattices, counting problems can often be solved by a variety of different methods, for example, transfer matrices and Pfaffians, which require a number of operations which are polynomial in the number of vertices.

The complexity of the counting problems changes if one considers nonplanar graphs, that is, graphs with a nontrivial topological genus. In discrete mathematics, such problems are classified as #P-complete, meaning that the existence of an exact polynomial algorithm for the evaluation of the generating functions would imply the polynomial solvability of many known counting combinatorial problems, the most famous one being the evaluation of the permanent of 0–1 matrices. In statistical mechanics and mathematical chemistry, the interest in nonplanar lattices is obviously related to their $D > 2$ character: the three-dimensional cubic lattice is nothing but a nonplanar graph of topological genus $g = 1 + N/4$, where N is the number of sites.

The planar two-dimensional Ising model was solved in 1944 by Onsager using the algebraic transfer matrix method. Successively, alternative exact solutions have been proposed which resorted to simple combinatorial and geometrical reasoning. As is well known, the underlying idea of the combinatorial methods consists in recasting the sum over spin configurations of the Boltzmann weights as a sum over closed curves (loops) weighted by the activity of their bonds. Double counting is avoided by a proper cancellation mechanism which takes care of the different intrinsic topologies of loops which give rise to the same contribution in the partition function. Such an approach has been developed first by Kac and Ward (1952) and provides a direct way of taking the field theoretic continuum limit. In $D > 2$, the generalization of the above method encounters enormous difficulties due to the variety of intrinsic topologies of surfaces immersed in $D > 2$ lattices.

Another combinatorial method proposed in the 1960s by Kasteleyn is the so-called Pfaffian method. It consists in writing the weighted sum over loops as a dimer covering or perfect matching generating function. Once the relationship between loop counting and dimer coverings (or perfect matchings) over a suitably decorated and properly oriented lattice is established, the Pfaffian method turns out to be a simple technique for the derivation of exact solutions or for the definition of polynomial algorithms over planar lattices which are applicable also to the two-dimensional Ising spin glass.

The generalization of the Pfaffian construction to the nonplanar case must deal with the ambiguity of orienting the homology cycles of the graph. Such a problem can be formally solved in full generality for any orientable lattice and leads to an expression of the Ising partition function or the dimer coverings generating function given as a sum over all possible inequivalent orientations of the lattice (or its embedding surface): for a graph of genus g , the homology basis is composed of $2g$ cycles and, therefore, there are 2^{2g} inequivalent orientations. It is only for graphs of logarithmic genus that the generalized Pfaffian formalism provides a polynomial algorithm.

Counting perfect matchings can be thought of as the problem of evaluating the permanent of 0–1 matrices over properly constructed bipartite graphs, which is among the oldest and most famous #P-complete problems.

The Pfaffian formalism when applied to the permanent problem leads to a simple general result, that is, it provides a general formula for writing the permanent of a matrix in terms of a number of determinants which is exponential in the genus of the underlying graph.

See also: Combinatorics: Overview; Determinantal Random Fields; Dimer Problems; Phase Transitions in Continuous Systems; Spin Glasses; Two-Dimensional Ising Model.

Further Reading

- Achlioptas D, Naor A, and Peres Y (2005) Rigorous location of phase transitions in hard optimization problems. *Nature* 435: 759–764.
- Ajtai N (1996) Generating hard instances of lattice problems. *Electronic Colloquium on Computational Complexity (ECCC)* 7: 3.
- Braunstein A, Mézard M, and Zecchina R (2005) Survey Propagation: an Algorithm for Satisfiability. *Random Structures and Algorithms* 27: 201–226.
- Cocco S and Monasson R (2004) Heuristic average-case analysis of the backtrack resolution of random 3-satisfiability instances. *Theoretical Computer Science A* 320: 345.
- Distler J (1992) A note on the 3D Ising model as a string theory. *Nuclear Physics B* 388: 648.
- Dubois O, Monasson R, Selman B, and Zecchina R (eds.) (2001) *NP-hardness and Phase transitions* (Special Issue), *Theoretical Computer Science*, vol. 265 (1–2). Elsevier.
- Friedgut E (1999) Sharp threshold of graph properties, and the KSat problem. *Journal of American Mathematical Society* 12: 1017–1054.
- Jerrum M and Sinclair A (1989) Approximating the permanent. *SIAM Journal on Computing* 18: 1149.
- Lovász L and Plummer MD (1986) *Matching Theory*. *North-Holland Mathematics Studies* 121, *Annals of Discrete Mathematics* (29). New York: North-Holland.
- Mézard M, Mora T, and Zecchina R (2005) Clustering of solutions in the random satisfiability problem. *Physics Review Letters* 94: 197205.

- Mézard M and Parisi G (2001) The Bethe lattice spin glass revisited. *European Physical Journal B* 20: 217.
- Mézard M, Parisi G, and Virasoro MA (1987) *Sping Glass Theory and Beyond*. Singapore: World Scientific.
- Mézard M, Parisi G, and Zecchina R (2002) Analytic and algorithmic solution of random satisfiability problems. *Science* 297: 812–815.
- Mézard M and Zecchina R (2002) Random K-satisfiability: from an analytic solution to a new efficient algorithm. *Physical Review E* 66: 056126.
- Monasson R, Zecchina R, Kirkpatrick S, Selman B, and Troyansky L (1999) Determining computational complexity from characteristic “phase transitions”. *Nature* 400: 133.
- Motwani R and Raghavan P (2000) *Randomized Algorithms*. Cambridge: Cambridge University Press.
- Nishimori H (2001) *Statistical Physics of Spin Glasses and Information Processing*. Oxford: Oxford University Press.
- Papadimitriou CH (1994) *Computational Complexity*. Addison-Wesley.
- Regge T and Zecchina R (2000) Combinatorial and topological approach to the 3D Ising model. *Journal of Physics A: Mathematical and General* 33: 741.
- Richardson T and Urbanke R (2001) An introduction to the analysis of iterative coding systems. In: Marcus B and Rosenthal J (eds.) *Codes, Systems, and Graphical Models*. New York: Springer.

Statistical Mechanics of Interfaces

S Miracle-Solé, Centre de Physique Théorique,
CNRS, Marseille, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction

When a fluid is in contact with another fluid, or with a gas, a portion of the total free energy of the system is proportional to the area of the surface of contact, and to a coefficient, the surface tension, which is specific for each pair of substances. Equilibrium will accordingly be obtained when the free energy of the surfaces in contact is a minimum.

Suppose that we have a drop of some fluid, b , over a flat substrate, w , while both are exposed to air, a . We have then three different surfaces of contact, and the total free energy of the system consists of three parts, associated to these three surfaces. A drop of fluid b will exist provided its own two surface tensions exceed the surface tension between the substrate w and the air, that is, provided that

$$\tau^{wb} + \tau^{ba} > \tau^{wa}$$

If equality is attained, then a film of fluid b is formed, a situation which is known as perfect, or complete wetting (see Figure 1).

When one of the substances involved is anisotropic, such as a crystal, the contribution to the total free energy of each element of area depends on its orientation. The minimum surface free energy for a given volume then determines the ideal form of the crystal in equilibrium.

It is only in recent times that equilibrium crystals have been produced in the laboratory, first, in negative crystals (vapor bubbles) of organic substances. Most crystals grow under nonequilibrium

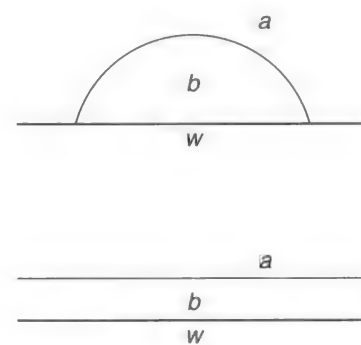


Figure 1 Partial and complete wetting.

conditions and it is a subsequent relaxation of the macroscopic crystal that restores the equilibrium.

An interesting phenomenon that can be observed on these crystals is the roughening transition, characterized by the disappearance of the facets of a given orientation, when the temperature attains a certain particular value. The best observations have been made on helium crystals, in equilibrium with superfluid helium, since the transport of matter and heat is then extremely fast. Crystals grow to sizes of 1–5 mm and relaxation times vary from milliseconds to minutes. Roughening transitions for three different types of facets have been observed (see, e.g., Wolf *et al.* (1983)).

These are some classical examples among a variety of interesting phenomena connected with the behavior of the interface between two phases in a physical system. The study of the nature and properties of the interfaces, at least for some simple systems in statistical mechanics, is also an interesting subject of mathematical physics. Some aspects of this study will be discussed in the present article.

We assume that the interatomic forces can be modeled by a lattice gas, and consider, as a simple example, the ferromagnetic Ising model. In a typical two-phase equilibrium state, there is a dense

component, which can be interpreted as a solid or liquid phase, and a dilute phase, which can be interpreted as the vapor phase. Considering certain particular cases of such situations, we first introduce a precise definition of the surface tension and then proceed on the mathematical analysis of some preliminary properties of the corresponding interfaces. The next topic concerns the wetting properties of the system, and the final section is devoted to the associated equilibrium crystal.

Pure Phases and Surface Tension

The Ising model is defined on the cubic lattice $\mathcal{L} = \mathbb{Z}^3$, with configuration space $\Omega = \{-1, 1\}^{\mathcal{L}}$. If $\sigma \in \Omega$, the value $\sigma(i) = -1$ or 1 is the spin at the site $i = (i_1, i_2, i_3) \in \mathcal{L}$, and corresponds to an empty or an occupied site in the lattice gas version of the model. The system is first considered in a finite box $\Lambda \subset \mathcal{L}$, with fixed values of the spins outside.

In order to simplify the exposition, we shall mainly consider the three-dimensional Ising model, though some of the results to be discussed hold in any dimension $d \geq 2$. We shall also, sometimes, refer to the two-dimensional model, it being understood that the definitions have been adapted in the obvious way. We assume that the box Λ is a parallelepiped, centered at the origin of \mathcal{L} , of sides L_1, L_2, L_3 , parallel to the axes.

A configuration of spins on $\Lambda(\sigma(i), i \in \Lambda)$, denoted σ_Λ , has an energy defined by the Hamiltonian

$$H_\Lambda(\sigma_\Lambda|\bar{\sigma}) = -J \sum_{\langle i,j \rangle \cap \Lambda \neq \emptyset} \sigma(i)\sigma(j) \quad [1]$$

where J is a positive constant (ferromagnetic or attractive interaction). The sum runs over all nearest-neighbor pairs $\langle i,j \rangle \subset \mathcal{L}$, such that at least one of the sites belongs to Λ , and one takes $\sigma(i) = \bar{\sigma}(i)$ when $i \notin \Lambda$, the configuration $\bar{\sigma} \in \Omega$ being the given boundary condition. The probability of the configuration σ_Λ , at the inverse temperature $\beta = 1/kT$, is given by the Gibbs measure

$$\mu_\Lambda(\sigma_\Lambda|\bar{\sigma}) = Z^\sigma(\Lambda)^{-1} \exp(-\beta H_\Lambda(\sigma_\Lambda|\bar{\sigma})) \quad [2]$$

where $Z^\sigma(\Lambda)$ is the partition function

$$Z^\sigma(\Lambda) = \sum_{\sigma_\Lambda} \exp(-\beta H_\Lambda(\sigma_\Lambda|\bar{\sigma})) \quad [3]$$

Local properties at equilibrium can be described by the correlation functions between the spins on finite sets of sites,

$$\mu_\Lambda^\sigma(\sigma(A)) = \sum_{\sigma_\Lambda} \mu_\Lambda(\sigma_\Lambda|\bar{\sigma}) \prod_{i \in A} \sigma(i) \quad [4]$$

The measures [2] determine (by the Dobrushin–Lanford–Ruelle equations) the set of Gibbs states of the infinite system, as measures on the set Ω of all configurations. If a Gibbs state happens to be equal to $\lim \mu_\Lambda(\cdot|\bar{\sigma})$, when $L_1, L_2, L_3 \rightarrow \infty$, under a fixed boundary condition $\bar{\sigma}$, we shall call it the Gibbs state associated to the boundary condition $\bar{\sigma}$. One also says that this state exists in the thermodynamic limit. Then, equivalently, the correlation functions [4] converge to the corresponding expectation values in this state.

This model presents, at low temperatures (i.e., for $\beta > \beta_c$, where β_c is the critical inverse temperature), two different thermodynamic pure phases, a dense and a dilute phase in the lattice gas language (called here the positive and the negative phase). This means two extremal translation-invariant Gibbs states, μ^+ and μ^- , obtained as the Gibbs states associated with the boundary conditions $\bar{\sigma}$, respectively equal to the ground configurations $\bar{\sigma}(i) = 1$ and $\bar{\sigma}(i) = -1$, for all $i \in \mathcal{L}$. The spontaneous magnetization

$$m^*(\beta) = \mu^+(\sigma(i)) = -\mu^-(\sigma(i)) \quad [5]$$

is then strictly positive. On the other hand, if $\beta \leq \beta_c$, then the Gibbs state is unique and $m^* = 0$.

Each configuration inside Λ can be described in a geometric way by specifying the set of Peierls contours which indicate the boundaries between the regions of spin 1 and the regions of spin -1 . Unit-square faces are placed midway between the pairs of nearest-neighbor sites i and j , perpendicularly to these bonds, whenever $\sigma(i)\sigma(j) = -1$. The connected components of this set of faces are the Peierls contours. Under the boundary conditions $(+)$ and $(-)$, the contours form a set of closed surfaces. They describe the defects of the considered configuration with respect to the ground states of the system (the constant configurations 1 and -1), and are a basic tool for the investigation of the model at low temperatures.

In order to study the interface between the two pure phases, one needs to construct a state describing the coexistence of these phases. This can be done by means of a new boundary condition. Let $\mathbf{n} = (n_1, n_2, n_3)$ be a unit vector in \mathbb{R}^3 , such that $n_3 > 0$, and introduce the mixed boundary condition (\pm, \mathbf{n}) , for which

$$\bar{\sigma}(i) = \begin{cases} 1 & \text{if } i \cdot \mathbf{n} \geq 0 \\ -1 & \text{if } i \cdot \mathbf{n} < 0 \end{cases} \quad [6]$$

This boundary condition forces the system to produce a defect going transversally through the box Λ , a large Peierls contour that can be

interpreted as the microscopic interface (also called a domain wall). The other defects that appear above and below the interface can be described by closed contours inside the pure phases.

The free energy per unit area due to the presence of the interface is the surface tension. It is defined by

$$\tau(\mathbf{n}) = \lim_{L_1, L_2 \rightarrow \infty} \lim_{L_3 \rightarrow \infty} -\frac{n_3}{\beta L_1 L_2} \ln \frac{Z^{\pm, \mathbf{n}}(\Lambda)}{Z^+(\Lambda)} \quad [7]$$

In this expression the volume contributions proportional to the free energy of the coexisting phases, as well as the boundary effects, cancel, and only the contributions to the free energy due to the interface are left. The existence of such a quantity indicates that the macroscopic interface, separating the regions occupied by the pure phases in a large volume Λ , has a microscopic thickness and can therefore be regarded as a surface in a thermodynamic approach.

Theorem 1 *The interfacial free energy per unit area, $\tau(\mathbf{n})$, exists, is bounded, and its extension by positive homogeneity, $f(\mathbf{x}) = |\mathbf{x}| \tau(\mathbf{x}/|\mathbf{x}|)$, is a convex function on \mathbb{R}^3 . Moreover, $\tau(\mathbf{n})$ is strictly positive for $\beta > \beta_c$, and vanishes if $\beta \leq \beta_c$.*

The existence of $\tau(\mathbf{n})$ and also the last statement were proved by Lebowitz and Pfister (1981), in the particular case $\mathbf{n} = (0, 0, 1)$, with the help of correlation inequalities. A complete proof of the theorem was given later with similar arguments. The convexity of f is equivalent to the fact that the surface tension τ satisfies a thermodynamic stability condition known as the pyramidal inequality (see Messager *et al.* (1992)).

Gibbs States and Interfaces

In this section we consider the (\pm, \mathbf{n}_0) boundary condition, also simply denoted (\pm) , associated to the vertical direction $\mathbf{n}_0 = (0, 0, 1)$,

$$\bar{\sigma}(i) = 1 \text{ if } i_3 \geq 0, \quad \bar{\sigma}(i) = -1 \text{ if } i_3 < 0 \quad [8]$$

The corresponding surface tension is $\tau^\pm = \tau(\mathbf{n}_0)$. We shall first recall some classical results which concern the Gibbs states and interfaces at low temperatures.

According to the geometrical description of the configurations introduced in the last section, we observe that

$$Z^{\pm, \mathbf{n}}(\Lambda)/Z^+(\Lambda) = \sum_{\lambda} \exp(-2\beta J|\lambda| - U_{\Lambda}(\lambda)) \quad [9]$$

where the sum runs over all microscopic interfaces λ compatible with the boundary condition and $|\lambda|$ is

the number of faces of λ (inside Λ). The term $U_{\Lambda}(\lambda)$ equals $-\ln Z^+(\lambda, \Lambda)/Z^+(\Lambda)$, the sum in the partition function $Z^+(\lambda, \Lambda)$ being extended to all configurations whose associated contours do not intersect λ . Each term in sum [9] gives a weight proportional to the probability of the corresponding microscopic interface.

At low (positive) temperatures, we expect the microscopic interface corresponding to this boundary condition, which at zero temperature coincides with the plane $i_3 = -1/2$, to be modified by small deformations. Each microscopic interface λ can then be described by its defects, with respect to the interface at $\beta = \infty$. To this end, one introduces some objects, called walls, which form the boundaries between the horizontal plane portions of the microscopic interface, also called the ceilings of the interface.

More precisely, one says that a face of λ is a ceiling face if it is horizontal and such that the vertical line passing through its center does not have other intersections with λ . Otherwise, one says that it is a wall face. The set of wall faces splits into maximal connected components. The set of walls, associated to λ , is the set of these components, each component being identified by its geometric form and its projection on the plane $i_3 = -1/2$. Every wall ω , with projection $\pi(\omega)$, increases the energy of the interface by a quantity $2J\|\omega\|$, where $\|\omega\| = |\omega| - |\pi(\omega)|$, and two walls are compatible if their projections do not intersect. In this way, the microscopic interfaces may be interpreted as a “gas of walls” on the two-dimensional lattice.

Dobrushin, who developed the above analysis, also proved the dilute character of this “gas” at low temperatures. This implies that the microscopic interface is essentially flat, or rigid. One can understand this fact by noticing first that the probability of a wall is less than $\exp(-2\beta J\|\omega\|)$ and, second, that in order to create a ceiling in λ , which is not in the plane $i_3 = -1/2$, one needs to surround it by a wall, that one has to grow when the ceiling is made over a larger area.

Using correlation inequalities one proves that the Gibbs state μ^\pm , associated to the (\pm) boundary conditions, always exists, and that it is invariant under horizontal translations of the lattice, that is, $\mu^\pm(\sigma(A+a)) = \mu^\pm(\sigma(A))$ for all $a = (a_1, a_2, 0)$. It is also an extremal Gibbs state. Let $m(z)$ be the magnetization $\mu^\pm(\langle \sigma(z) \rangle)$ at the site $z = (0, 0, z)$. The function $m(z)$ is monotone increasing and satisfies the symmetry property $m(-z) = -m(z+1)$. Some consequences of Dobrushin’s work are the following properties.

Theorem 2 *If the temperature is low enough, that is, if $\beta J \geq c_1$, where c_1 is a given constant, then*

$m^\pm(0)$ is strictly positive [10]

$m^\pm(z) \rightarrow m^*$, when $z \rightarrow \infty$, exponentially fast [11]

Equation [10] is just another way of saying that the interface is rigid and that the state μ^\pm is non-translation invariant (in the vertical direction). Then, the correlation functions $\mu^\pm(\sigma(A))$ describe the local properties, or local structure, of the macroscopic interface. In particular, the function $m(z)$ represents the magnetization profile. Then statement [11], together with the symmetry property, tells us that the thickness of this interface is finite, with respect to the unit lattice spacing.

The statistics of interfaces has been rewritten in terms of a gas of walls and this system may further be studied by cluster expansion techniques. There is an interaction between the walls, coming from the term $U_\Lambda(\lambda)$ in eqn [9], but a convenient mathematical description of this interaction can be obtained by applying the standard low-temperature cluster expansion, in terms of contours, to the regions above and below the interface.

This method was introduced by Gallavotti in his study (mentioned below) of the two-dimensional Ising model. It has been applied by Bricmont and co-workers to examine the interface structure in the present case. As a consequence, it follows that the surface tension, more exactly $\beta\tau^\pm(\beta)$, and also the correlation functions, are analytic functions at low temperatures. They can be obtained as explicit convergent series in the variable $\zeta = e^{-2\beta J}$.

The same analysis applied to the two-dimensional model shows a very different behavior at low temperatures. In this case, the microscopic interface λ is a polygonal line and the walls belong to the one-dimensional lattice. One can then increase the size of a ceiling without modifying the walls attached to it.

Indeed, Gallavotti turned this observation into a proof that the Gibbs state μ^\pm is now translation invariant. The line λ undergoes large fluctuations of order $\sqrt{L_1}$, and disappears from any finite region of the lattice, in the thermodynamic limit. In particular, we have then $\mu^\pm = (1/2)(\mu^+ + \mu^-)$, a result that extends to all boundary conditions (\pm, n) .

Using these results Bricmont and co-workers also studied the local structure of the interface at low temperatures and showed that its intrinsic thickness is finite. To study the global fluctuations, one can compute the magnetization profile by introducing, before taking the thermodynamic limit, a change of scale: $\mu^\pm_\delta(\sigma(zL_1^\delta))$, with $\delta = 1/2$ or near to this value.

This is an exact computation that has been done by Abraham and Reed.

Let us come back to the three-dimensional Ising model where we know that the interface orthogonal to a lattice axis is rigid at low temperatures.

Question 1 At higher temperatures, but before reaching the critical temperature, do the fluctuations of this interface become unbounded, in the thermodynamic limit, so that the corresponding Gibbs state is translation invariant?

One says then that the interface is rough, and it is believed that, effectively, the interface becomes rough when the temperature is raised, undergoing a roughening transition at an inverse temperature $\beta_R > \beta_c$.

It is known that $\beta_R \leq \beta_c^{d=2}$, the critical inverse temperature of the two-dimensional Ising model, since van Beijeren proved, using correlation inequalities, that above this value, the state μ^\pm is not translation invariant. Recalling that the rigid interface may be viewed as a two-dimensional system, the system of walls, a representation that would become inappropriate for a rough interface, one might think that the phase transition of the two-dimensional Ising model is relevant for the roughening transition, and that β_R is somewhere near $\beta_c^{d=2}$. Indeed, approximate methods, used by Weeks and co-workers give some evidence for the existence of such a β_R and suggest a value slightly smaller than $\beta_c^{d=2}$, as shown in Table 1. To this day, however, there appears to be no proof of the fact that $\beta_R > \beta_c$, that is, that the roughening transition for the three-dimensional Ising model really occurs.

At present one is able to study the roughening transition rigorously only for some simplified models with a restricted set of admissible microscopic interfaces. Moreover, the closed contours, describing the defects above and below λ , are neglected, so that these two regions have the constant configurations 1 or -1, and one has $U_\Lambda(\lambda) = 0$ in eqn [9].

The best known of these models is the classic SOS (solid-on-solid) model in which the interfaces λ have the property of being cut only once by all vertical lines of the lattice. This means that λ is the graph of a function that can equivalently be used to define the possible configurations of λ . If λ contains the horizontal face with center $(i_1, i_2, i_3 - 1/2)$, then

Table 1 Some temperature values

$d=3$	$\beta_c J \sim 0.22$	approximate critical temperature
$d=3$	$\beta_R J \sim 0.41$	conjectured roughening temperature
$d=2$	$\beta_c J = 0.44$	exact critical temperature

the value at (i_1, i_2) of the associated function is $\phi(i_1, i_2) = i_3$.

The proof that the SOS model with the boundary condition (\pm) has a roughening transition is a highly nontrivial result due to Fröhlich and Spencer. When β is small enough, the fluctuations of λ are of order $\sqrt{\ln L}$ (in a cubic box of side L).

Moreover, other interface models, with additional conditions on the allowed microscopic interfaces, are exactly solvable. The BCSOS (body-centered SOS) model, introduced by van Beijeren, belongs to this class. It is, in fact, the first model for which the existence of a roughening transition has been proved. More recently, also the TISOS (triangular Ising SOS) model, introduced by Blöte and Hilhorst and further studied by Nienhuis and co-workers, has been considered in this context.

The interested reader can find more information and references, concerning the subject of this section, in the review article by Abraham (1986).

Wetting Phenomena

Next we consider the Ising model over a plane horizontal substrate (also called a wall) and study the difference of surface tensions which governs the wetting properties of this system.

We first describe the approach developed by Fröhlich and Pfister (1987) and briefly report some results of their study. We consider the model on the semi-infinite lattice

$$\mathcal{L}' = \{i \in \mathbb{Z}^3 : i_3 \geq 0\} \quad [12]$$

A magnetic field, $K \geq 0$, is added on the boundary sites, $i_3 = 0$, which describes the interaction with the substrate, supposed to occupy the complementary region $\mathcal{L} \setminus \mathcal{L}'$.

We constrain the model in the finite box $\Lambda' = \Lambda \cap \mathcal{L}'$, with Λ as above, and impose the value of the spins outside. The Hamiltonian becomes

$$H_{\Lambda'}^w(\sigma_{\Lambda'} | \bar{\sigma}) = -J \sum_{\langle i, j \rangle \cap \Lambda' \neq \emptyset} \sigma(i) \sigma(j) - K \sum_{i \in \Lambda', i_3 = 0} \sigma(i) \quad [13]$$

Here $\sigma_{\Lambda'}$ represents the configuration inside Λ' , the pairs $\langle i, j \rangle$ are contained in \mathcal{L}' , and $\sigma(i) = \bar{\sigma}(i)$ when $i \notin \Lambda'$, the configuration $\bar{\sigma}$ being the given boundary condition (see Figure 2). The corresponding partition function is denoted by $Z^{w\bar{\sigma}}(\Lambda')$.

Since there are two pure phases in the model, we must consider two surface free energies, or surface tensions, τ^{w+} and τ^{w-} , between the wall and the positive or negative phase present in the bulk. They are defined through the choice of the boundary

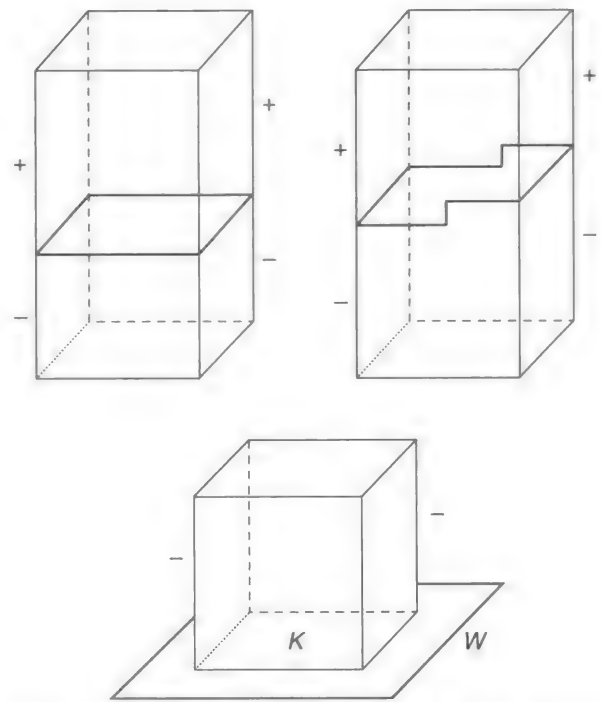


Figure 2 Boundary conditions for the cubic lattice. Above, the box Λ with the (\pm) and (step) boundary conditions. Below, the box Λ' and the wall W with the $(w-)$ boundary conditions.

condition, $\bar{\sigma}(i) = 1$ or $\bar{\sigma}(i) = -1$, for all $i \in \mathcal{L}'$. Let us consider first the case of the $(-)$ boundary condition.

The surface free energy contribution per unit area due to the presence of the wall, when we have the negative phase in the bulk, is

$$\tau^{w-}(\beta, K) = \lim_{L_1, L_2 \rightarrow \infty} \lim_{L_3 \rightarrow \infty} -\frac{1}{\beta L_1 L_2} \ln \frac{Z^{w-}(\Lambda')}{Z^{-}(\Lambda)^{1/2}} \quad [14]$$

The division by $Z^{-}(\Lambda)^{1/2}$ allows us to subtract from the total free energy, $\ln Z^{w-}(\Lambda')$, the bulk term and all boundary terms which are not related to the presence of the wall. The existence of limit [14] follows from correlation inequalities, and we have $\tau^{w-} \geq 0$.

One can prove, as well, the existence of the Gibbs state μ^{w-} of the semi-infinite system, associated to the $(-)$ boundary condition. This state is the limit of the finite volume Gibbs measures $\mu_{\Lambda'}(\sigma_{\Lambda'} | (-))$ defined by the Hamiltonian [13]. It describes the local equilibrium properties of the system near the wall, when deep inside the bulk the system is in the negative phase. Similar definitions give the surface tension τ^{w+} and the Gibbs state μ^{w+} , corresponding to the boundary condition $\bar{\sigma}(i) = 1$, for all $i \in \Lambda'$.

We remark that the states μ^{w+} and μ^{w-} are invariant by translations parallel to the plane $i_3 = 0$, and introduce the magnetizations, $m^{w-}(z) = \mu^{w-}(\sigma(z))$, where z denotes the site $(0, 0, z)$, $m^{w-} = m^{w-}(0)$, and

similarly $m^{w+}(z)$ and m^{w+} . Their connection with the surface free energies is given by the formula

$$\begin{aligned} \tau^{w-}(\beta, K) - \tau^{w+}(\beta, K) \\ = \int_0^K (m^{w+}(\beta, s) - m^{w-}(\beta, s)) ds \end{aligned} \quad [15]$$

We mention in the following theorem some results of Fröhlich and Pfister's study. Here τ^\pm is, as before, the usual surface tension between the two pure phases of the system, for a horizontal interface.

Theorem 3 *With the above definitions, we have*

$$\tau^{w-}(\beta, K) - \tau^{w+}(\beta, K) \leq \tau^\pm(\beta) \quad [16]$$

$$m^{w+}(\beta, K) - m^{w-}(\beta, K) \geq 0 \quad [17]$$

and the difference in [17] is a monotone decreasing function of the parameter K . Moreover, if $m^{w+} = m^{w-}$, then the Gibbs states μ^{w+} and μ^{w-} coincide.

The proof is a subtle application of correlation inequalities. Since, from Theorem 3, the integrand in eqn [15] is a positive and decreasing function, the difference $\Delta\tau = \tau^{w-} - \tau^{w+}$ is a monotone increasing and concave (and hence continuous) function of the parameter K . On the other hand, one can prove that $\Delta\tau = \tau^\pm$, if $K \geq J$. This justifies the following definition:

$$K_w(\beta) = \min\{K: \Delta\tau(\beta, K) = \tau^\pm(\beta)\} \quad [18]$$

In the thermodynamic description of wetting, the partial-wetting regime is characterized by the strict inequality in [16]. Equivalently, by $K < K_w(\beta)$. We must have then $m^{w+} \neq m^{w-}$, because of eqn [15]. This shows that, in the case of partial wetting, μ^{w+} and μ^{w-} are different Gibbs states.

The complete-wetting regime is characterized by the equality in [16], that is, by $K \geq K_w(\beta)$. Then, we have $m^{w+} = m^{w-}$, and taking into account the last statement in Theorem 3, also $\mu^{w+} = \mu^{w-}$. This last result implies that there is only one Gibbs state. Thus, complete wetting corresponds to unicity of the Gibbs state.

In this case, we also have $\lim_{z \rightarrow \infty} m^{w-}(z) = m^*$, when $z \rightarrow \infty$, because this is always true for $m^{w+}(z)$. This indicates that we are in the positive phase of the system although we have used the $(-)$ boundary condition, so that the bulk negative phase cannot reach the wall anymore. The film of positive phase, which wets the wall completely, has an infinite thickness with respect to the unit lattice spacing, in the thermodynamic limit.

When $\beta = \infty$, only a few particular ground configurations contribute to the partition functions, such as the configuration $\sigma(i) = -1$ for the partition

function Z^{w-} , etc., and we obtain $\Delta\tau = 2K$ and $\tau^\pm = 2J$. For nonzero but low temperatures, the small perturbations of these ground states have to be considered, a problem that can be treated by the method of cluster expansions. In fact, the corresponding defects can be described by closed contours as in the case of pure phases.

Theorem 4 *For $K < J$, the functions $\beta\tau^{w-}(\beta, K)$ and $\beta\tau^{w+}(\beta, K)$ are analytic at low temperatures, that is, provided that $\beta(J - K) \geq c_2$, where c_2 is a given constant. Moreover, $m^{w+}(z)$ and $m^{w-}(z)$ tend, respectively, to m^* and to $-m^*$, when $z \rightarrow \infty$, exponentially fast.*

The last statement in Theorem 4 tells us that the wall affects only a layer of finite thickness (with respect to the lattice spacing). From a macroscopic point of view, the negative phase reaches the wall, and we are in the partial-wetting regime. Indeed, a strict inequality holds in [16].

Thus, for $K < J$ there is always partial wetting at low temperatures. Then the following question arises:

Question 2 Is there a situation of complete wetting at higher temperatures? It is understood here that K takes a fixed value, characteristic of the substrate, such that $0 < K < J$.

This is known to be the case in dimension $d = 2$, where the exact value of $K_w(\beta)$ can be obtained from Abraham's solution of the model:

$$\cosh 2\beta K_w = \cosh 2\beta J - e^{-2\beta J} \sinh 2\beta J$$

Then complete wetting occurs for β in the interval $\beta_c < \beta \leq \beta_w(K)$, where β_c is the critical inverse temperature and $\beta_w(K)$ is the solution of $K_w(\beta) = K$. The case $d = 2$ has been reviewed in Abraham (1986).

To our knowledge, the above question remains an open problem for the Ising model in dimension $d = 3$. The problem has, however, been solved for the simpler case of a SOS interface model. In this case, a nice and rather brief proof of the following result has been given by Chalker (1982): one has $m^{w+} = m^{w-}$, and hence complete wetting, if

$$2\beta(J - K) < -\ln(1 - e^{-8\beta J})$$

It is very plausible that a similar statement is valid for the semi-infinite Ising model and, also that Chalker's method could play a role for extending the proof to this case, provided an additional assumption is made. Namely, that β is sufficiently large, and hence $J - K$ small enough, in order to insure the convergence of the cluster expansions and to be able to use them.

Equilibrium Crystals

The shape of an equilibrium crystal is obtained, according to thermodynamics, by minimizing the surface free energy between the crystal and the medium, for a fixed volume of the crystal phase. Given the orientation-dependent surface tension $\tau(\mathbf{n})$, the solution to this variational problem, known under the name of Wulff construction, is the following set:

$$\mathcal{W} = \{\mathbf{x} \in \mathbb{R}^3: \mathbf{x} \cdot \mathbf{n} \leq \tau(\mathbf{n}) \text{ for all } \mathbf{n}\} \quad [19]$$

Notice that the problem is scale invariant, so that if we solve it for a given volume of the crystal, we get the solution for other volumes by an appropriate scaling. We notice also that the symmetry $\tau(\mathbf{n}) = \tau(-\mathbf{n})$ is not required for the validity of formula [19]. In the present case, $\tau(\mathbf{n})$ is obviously a symmetric function, but nonsymmetric situations are also physically interesting and appear, for instance, in the case of a drop on a wall discussed in the last section.

The surface tension in the Ising model between the positive and negative phases has been defined in eqn [7]. In the two-dimensional case, this function $\tau(\mathbf{n})$ has (as shown by Abraham) an exact expression in terms of some Onsager's function. It follows (as explained in Miracle-Sole (1999)) that the Wulff shape \mathcal{W} , in the plane (x_1, x_2) , is given by

$$\cosh \beta x_1 + \cosh \beta x_2 \leq \cosh^2 2\beta J / \sinh 2\beta J$$

This shape reduces to the empty set for $\beta \leq \beta_c$, since the critical β_c satisfies $\sinh 2\beta_c = 1$. For $\beta > \beta_c$, it is a strictly convex set with smooth boundary.

In the three-dimensional case, only certain interface models can be exactly solved (see the section "Gibbs states and interfaces"). Consider the Ising model at zero temperature. The ground configurations have only one defect, the microscopic interface λ , imposed by the boundary condition (\pm, \mathbf{n}) . Then, from eqn [9], we may write

$$\tau(\mathbf{n}) = \lim_{L_1, L_2 \rightarrow \infty} \frac{n_3}{L_1 L_2} (E_\lambda(\mathbf{n}) - \beta^{-1} N_\lambda(\mathbf{n})) \quad [20]$$

where $E_\lambda = 2J|\lambda|$ is the energy (all λ have the same minimal area) and N_λ the number of ground states. Every such λ has the property of being cut only once by all straight lines orthogonal to the diagonal plane $i_1 + i_2 + i_3 = 0$, provided that $n_k > 0$, for $k=1, 2, 3$. Each λ can then be described by an integer function defined on a triangular plane lattice, the projection of the cubic lattice \mathcal{L} on the diagonal plane. The model defined by this set of admissible microscopic

interfaces is precisely the TISOS model. A similar definition can be given for the BCSOS model that describes the ground configurations on the body-centered cubic lattice.

From a macroscopic point of view, the roughness or the rigidity of an interface should be apparent when considering the shape of the equilibrium crystal associated with the system. A typical equilibrium crystal at low temperatures has smooth plane facets linked by rounded edges and corners. The area of a particular facet decreases as the temperature is raised and the facet finally disappears at a temperature characteristic of its orientation. It can be argued that the disappearance of the facet corresponds to the roughening transition of the interface whose orientation is the same as that of the considered facet.

The exactly solvable interface models mentioned above, for which the function $\tau(\mathbf{n})$ has been computed, are interesting examples of this behavior, and provide a valuable information on several aspects of the roughening transition. This subject has been reviewed by Abraham (1986), van Beijeren and Nolden (1987), and Kotecky (1989).

For example, we show in Figure 3 the shape predicted by the TISOS model (one-eighth of the shape because of the condition $n_k > 0$). In this model, the interfaces orthogonal to the three coordinate axes are rigid at low temperatures.

For the three-dimensional Ising model at positive temperatures, the description of the microscopic interface, for any orientation \mathbf{n} , appears as a very difficult problem. It has been possible, however, to analyze the interfaces which are very near to the particular orientations \mathbf{n}_0 , discussed in the

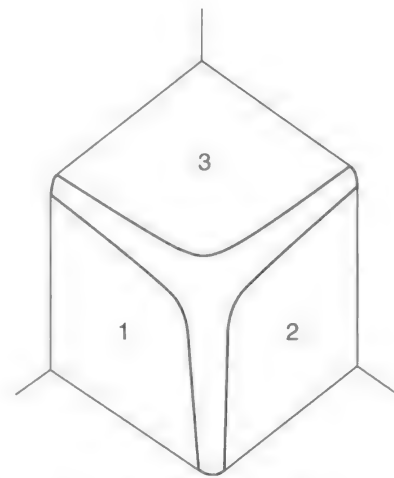


Figure 3 Cubic equilibrium crystal shown in a projection parallel to the (1,1,1) direction. The three regions (1, 2, and 3) indicate the facets and the remaining area represents a curved part of the crystal surface.

section “Gibbs states and interfaces.” This analysis allows us to determine the shape of the facets in a rigorous way.

We first observe that the appearance of a facet in the equilibrium crystal shape is related, according to the Wulff construction, to the existence of a discontinuity in the derivative of the surface tension with respect to the orientation. More precisely, assume that the surface tension satisfies the convexity condition of Theorem 1, and let this function $\tau(\mathbf{n}) = \tau(\theta, \phi)$ be expressed in terms of the spherical coordinates of \mathbf{n} , the vector \mathbf{n}_0 being taken as the x_3 -axis. A facet orthogonal to \mathbf{n}_0 appears in the Wulff shape if and only if the derivative $\partial\tau(\theta, \phi)/\partial\theta$ is discontinuous at the point $\theta=0$, for all ϕ . The facet $\mathcal{F} \subset \partial\mathcal{W}$ consists of the points $\mathbf{x} \in \mathbb{R}^3$ belonging to the plane $x_3 = \tau(\mathbf{n}_0)$ and such that, for all ϕ between 0 and 2π ,

$$x_1 \cos \phi + x_2 \sin \phi \leq \partial\tau(\theta, \phi)/\partial\theta|_{\theta=0^+} \quad [21]$$

The step free energy is expected to play an important role in the facet formation. It is defined as the free energy associated with the introduction of a step of height 1 on the interface, and can be regarded as an order parameter for the roughening transition. Let Λ be a parallelepiped as in the section “Pure phases and surface tension,” and introduce the (step, \mathbf{m}) boundary conditions (see Figure 2), associated to the unit vectors $\mathbf{m} = (\cos \phi, \sin \phi) \in \mathbb{R}^2$, by

$$\bar{\sigma}(i) = \begin{cases} 1 & \text{if } i > 0 \text{ or if } i_3 = 0 \text{ and} \\ & i_1 m_1 + i_2 m_2 \geq 0 \\ -1 & \text{otherwise} \end{cases} \quad [22]$$

Then, the step free energy per unit length for a step orthogonal to \mathbf{m} (with $m_2 > 0$) on the horizontal interface, is

$$\begin{aligned} \tau^{\text{step}}(\phi) \\ = \lim_{L_1 \rightarrow \infty} \lim_{L_2 \rightarrow \infty} \lim_{L_3 \rightarrow \infty} -\frac{\cos \phi}{\beta L_1} \ln \frac{Z^{\text{step}, \mathbf{m}}(\Lambda)}{Z^{\pm, \mathbf{n}_0}(\Lambda)} \end{aligned} \quad [23]$$

A first result concerning this point was obtained by Bricmont and co-workers, by proving a correlation inequality which establish $\tau^{\text{step}}(0)$ as a lower bound to the one-sided derivative $\partial\tau(\theta, 0)/\partial\theta$ at $\theta=0^+$ (the inequality extends also to $\phi \neq 0$). Thus, when $\tau^{\text{step}} > 0$, a facet is expected.

Using the perturbation theory of the horizontal interface, it is possible to also study the microscopic interfaces associated with the (step, \mathbf{m}) boundary conditions. When considering these configurations,

the step may be viewed as an additional defect on the rigid interface described in the section “Pure phases and surface tension.” It is, in fact, a long wall going from one side to the other side of the box Λ . The step structure at low temperatures can then be analyzed with the help of a new cluster expansion. As a consequence of this analysis, we have the following theorem.

Theorem 5 *If the temperature is low enough, that is, if $\beta J \geq c_3$, where c_3 is a given constant, then the step free energy, $\tau^{\text{step}}(\phi)$, exists, is strictly positive, and extends by positive homogeneity to a strictly convex function. Moreover, $\beta\tau^{\text{step}}(\phi)$ is an analytic function of $\zeta = e^{-2J\beta}$, for which an explicit convergent series expansion can be found.*

Using the above results on the step structure, similar methods allow us to evaluate the increment in surface tension of an interface tilted by a very small angle θ with respect to the rigid horizontal interface. This increment can be expressed in terms of the step free energy, and one obtains the following relation.

Theorem 6 *For $\beta J \geq c_3$, we have*

$$\partial\tau(\theta, \phi)/\partial\theta|_{\theta=0^+} = \tau^{\text{step}}(\phi) \quad [24]$$

This relation, together with eqn [21], implies that one obtains the shape of the facet by means of the two-dimensional Wulff construction applied to the step free energy. The reader will find a detailed discussion on these points, as well as the proofs of Theorems 5 and 6, in Miracle-Sole (1995).

From the properties of τ^{step} stated in Theorem 5, it follows that the Wulff equilibrium crystal presents well-defined boundary lines, smooth and without straight segments, between a rounded part of the crystal surface and the facets parallel to the three main lattice planes.

It is expected, but not proved, that at a higher temperature, but before reaching the critical temperature, the facets associated with the Ising model undergo a roughening transition. It is then natural to believe that the equality [24] is true for any β larger than β_R , allowing us to determine the facet shape from eqns [21] and [24], and that for $\beta \leq \beta_R$, both sides in this equality vanish, and thus, the disappearance of the facet is involved. However, the condition that the temperature is low enough is needed in the proofs of Theorems 5 and 6.

See also: Dimer Problems; Phase Transitions in Continuous Systems; Phase Transition Dynamics; Two-Dimensional Ising Model; Wulff Droplets.

Further Reading

- Abraham DB (1986) Surface structures and phase transitions. In: Domb C and Lebowitz JL (eds.) *Critical Phenomena*, vol. 10, pp. 1–74. London: Academic Press.
- Chalker JT (1982) The pinning of an interface by a planar defect. *Journal of Physics A: Mathematical and General* 15: L481–L485.
- Fröhlich J and Pfister CE (1987) Semi-infinite Ising model. I. *Communications in Mathematical Physics* 109: 493–523.
- Fröhlich J and Pfister CE (1987) Semi-infinite Ising model. II. *Communication in Mathematical Physics* 112: 51–74.
- Gallavotti G (1999) *Statistical Mechanics: A Short Treatise*. Berlin: Springer.
- Kotecky R (1989) Statistical mechanics of interfaces and equilibrium crystal shapes. In: Simon B, Truman A, and Davies IM (eds.) *IX International Congress of Mathematical Physics*, pp. 148–163. Bristol: Adam Hilger.
- Lebowitz JL and Pfister CE (1981) Surface tension and phase coexistence. *Physical Review Letters* 46: 1031–1033.

- Messenger A, Miracle-Sole S, and Ruiz J (1992) Convexity properties of the surface tension and equilibrium crystals. *Journal of Statistical Physics* 67: 449–470.
- Miracle-Sole S (1995) Surface tension, step free energy and facets in the equilibrium crystal shape. *Journal of Statistical Physics* 79: 183–214.
- Miracle-Sole S (1999) Facet shapes in a Wulff crystal. In: Miracle-Sole S, Ruiz J, and Zagrebnov V (eds.) *Mathematical Results in Statistical Mechanics*, pp. 83–101. Singapore: World Scientific.
- van Beijeren H and Nolden I (1987) The roughening transition. In: Schommers W and von Blackenhagen P (eds.) *Topics in Current Physics*, vol. 43, pp. 259–300. Berlin: Springer.
- Widom B (1991) Interfacial phenomena. In: Hansen JP, Levesque D, and Zinn-Justin J (eds.) *Liquid, Freezing and the Glass Transition*, pp. 506–546. Amsterdam: North-Holland.
- Wolf PE, Balibar S, and Gallet F (1983) Experimental observations of a third roughening transition in hcp ^4He crystals. *Physical Review Letters* 51: 1366–1369.

Stochastic Differential Equations

F Russo, Université Paris 13, Villetaneuse, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Stochastic differential equations (SDEs) appear today as a modeling tool in several sciences as telecommunications, economics, finance, biology, and quantum field theory.

An SDE is essentially a classical differential equation which is perturbed by a random noise. When nothing else is specified, SDE means in fact ordinary SDE; in that case it corresponds to the perturbation of an ordinary differential equation. Stochastic partial differential equations (SPDEs) are obtained as random perturbation of partial differential equations (PDEs).

One of the most important difference between deterministic and stochastic ordinary differential equations is described by the so-called Peano type phenomenon. A classical differential equation with continuous and linear growth coefficients admits global existence but not uniqueness as classical calculus text books illustrate studying equations of the type

$$\frac{dX}{dt}(t) = \sqrt{X(t)}, \quad X(0) = 0$$

However, if one perturbs the right member of the equality with an additive Gaussian white noise (ξ_t) (even with very small intensity), then the problem

becomes well stated. A similar phenomenon happens with linear PDEs of evolution type perturbed with a spacetime white noise.

SDEs constitute a vast subject and account for an incredible amount of relevant contributions. We try to orientate the reader about the main axes trying to indicate references to the different subfields. We will prefer to refer to monographs when available, instead of articles.

Motivation and Preliminaries

In the whole article T will be a strictly positive real number. Let us consider continuous functions $b: \mathbb{R}_+ \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, $a: \mathbb{R}_+ \times \mathbb{R}^{d \times m} \rightarrow \mathbb{R}^d$ and $x_0 \in \mathbb{R}^d$. We consider a differential problem of the following type:

$$\begin{aligned} \frac{dX_t}{dt} &= b(t, X_t) \\ X_0 &= x_0 \end{aligned} \quad [1]$$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. Suppose that previous equation is perturbed by a random noise $(\xi_t)_{t \geq 0}$. Because of modeling reasons it could be reasonable to suppose $(\xi_t)_{t \geq 0}$ satisfying the following properties.

1. It is a family of independent random variables (r.v.'s)
2. $(\xi_t)_{t \geq 0}$ is “stationary”, that is, for any positive integer n , positive reals h, t_0, t_1, \dots, t_n , the law of $(\xi_{t_0+h}, \dots, \xi_{t_n+h})$ does not depend on h .

More precisely we perturb eqn [1] as follows:

$$\begin{aligned} \frac{dX_t}{dt} &= b(t, X_t) + a(t, X_t)\xi_t \\ X_0 &= x_0 \end{aligned} \quad [2]$$

We suppose for a moment that $d = m = 1$. In reality no reasonable real-valued process $(\xi_t)_{t \geq 0}$ fulfilling previous assumptions exists. In particular, if process (ξ_t) exists (resp. (ξ_t) exists and each ξ_t is a square-integrable r.v.), then the process cannot have continuous paths (resp. it cannot be measurable with respect to $\Omega \times \mathbb{R}_+$). However, suppose that such a process exists; we set $B_t = \int_0^t \xi_s ds$. In that case, properties (1) and (2) can be translated into the following on (B_t) .

- (P1) It has independent increments, which means that for any $t_0, \dots, t_n, h \geq 0, B_{t_1+h} - B_{t_0+h}, \dots, B_{t_n+h} - B_{t_{n-1}+h}$ are independent r.v.'s.
- (P2) It has stationary increments, which means that for any $t_0, \dots, t_n, h \geq 0$, the law of $(B_{t_1+h} - B_{t_0+h}, \dots, B_{t_n+h} - B_{t_{n-1}+h})$ does not depend on h .

On the other hand, it is natural to require that

- (C1) $B_0 = 0$ a.s.,
- (C2) it is a continuous process, that is, it has continuous paths a.s.

Equation [2] should be rewritten in some integral form

$$\begin{aligned} X_t &= X_0 + \int_0^t b(s, X_s) ds \\ &\quad + \int_0^t a(s, X_s) dB_s \end{aligned} \quad [3]$$

Clearly the paths of process (B_t) cannot be differentiable, so one has to give meaning to integral $\int_0^t a(s, X_s) dB_s$. This will be intended in the "Itô" sense, see considerations below.

An important result of probability theory says that a stochastic process (B_t) fulfilling properties P1, P2 and C1, C2 is essentially a "Brownian motion". More precisely, there are real constants b, σ such that $B_t = bt + \sigma W_t$, where (W_t) is a classical Brownian motion defined below.

Definition 1

- (i) A (continuous) stochastic process (W_t) is called classical "Brownian motion" if $W_0 = 0$ a.s., it has independent increments and the law of $W_t - W_s$ is a Gaussian $N(0, t - s)$ r.v.
- (ii) A m -dimensional Brownian motion is a vector (W^1, \dots, W^m) of independent classical Brownian motions.

Let $(\mathcal{F}_t)_{t \geq 0}$ be a filtration fulfilling the usual conditions, see (Karatzas and Shreve (1991, section 1.1).

There one can find basic concepts of the theory of stochastic processes as the concept of adapted, progressively measurable process. An adapted process is also said to be nonanticipating towards the filtration (\mathcal{F}_t) which represents the state of the information at each time t . A process (X_t) is said to be adapted if for any t, X_t is \mathcal{F}_t -measurable. The notion of progressively measurable process is a slight refinement of the notion of adapted process.

Definition 2

- (i) A (continuous) (\mathcal{F}_t) adapted process (W_t) is called (classical) (\mathcal{F}_t) -Brownian motion if $W_0 = 0$, if for any $s < t$ $W_t - W_s$ is an $N(0, t - s)$ distributed r.v. which is independent of \mathcal{F}_s .
- (ii) An (\mathcal{F}_t) - m -dimensional Brownian motion is a vector (W^1, \dots, W^m) of (\mathcal{F}_t) -classical independent Brownian motions.

From now on, we will consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a filtration $(\mathcal{F}_t)_{t \geq 0}$ fulfilling the usual conditions. From now on all the considered filtrations will have that property.

Let $W = (W_t)_{t \geq 0}$ be an $(\mathcal{F}_t)_{t \geq 0}$ - m -dimensional classical Brownian motion. In Karatzas and Shreve (1991, chapter 3) and Revuz and Yor (1999, chapter 4), one introduces the notion of stochastic Itô integral announced before. Let $Y = (Y^1, \dots, Y^m)$ be a progressively measurable m -dimensional process such that $\int_0^T \|Y_s\|^2 ds < \infty$, then the Itô integral $\int_0^T Y_s dW_s$ is well defined. In particular the indefinite integral $\int_0^t Y_s dW_s$ is an (\mathcal{F}_t) -progressively measurable continuous process. If Y is an $\mathbb{R}^{d \times m}$ matrix-valued process, the integral $\int_0^t Y_s dW_s$ is componentwise defined and it will be a vector in \mathbb{R}^d . The analogous of differential calculus in the framework of stochastic processes is Itô calculus, see again Karatzas and Shreve (1991, chapter 3) and Revuz and Yor (1999, chapter 4). Important tools are the concept of quadratic variation $[X]$ of a stochastic process when it exists. For instance, the quadratic variation $[W]_t$ of a classical Brownian motion equals t . If $M_t = \int_0^t Y_s dW_s$, then $[M]_t = \int_0^t \|Y_s\|^2 ds$. One celebrated theorem of P Lévy states the following: if (M_t) defines a continuous (\mathcal{F}_t) -local martingale such that $[M]_t \equiv t$, then M is an (\mathcal{F}_t) -classical Brownian motion. That theorem is called the "Lévy characterization theorem of Brownian motion." Itô formula constitutes the natural generalization of fundamental theorem of differential calculus to the stochastic calculus. Another significant tool is Girsanov theorem; it states essentially the following: suppose that the following so-called "Novikov condition" is verified:

$$\mathbb{E} \left(\exp \left(\frac{1}{2} \int_0^T \|Y_t\|^2 dt \right) \right) < \infty$$

Then the process $\tilde{W}_t = W_t + \int_0^t Y_s ds, t \in [0, T]$ is again an m -dimensional (\mathcal{F}_t) -classical Brownian motion under a new probability measure Q on (Ω, \mathcal{F}_T) defined by

$$dQ = dP \exp\left(\int_0^t Y_s dW_s - \frac{1}{2} \|Y_s\|^2 ds\right)$$

Let ξ be an \mathcal{F}_0 -measurable r.v., for instance, $\xi \equiv x \in \mathbb{R}^d$. We are interested in the SDE

$$\begin{aligned} dX_t &= a(t, X_t) dW_t + b(t, X_t) dt \\ X_0 &= \xi \end{aligned} \quad [4]$$

Definition 3 A progressively measurable process $(X_t)_{t \in [0, T]}$ is said to be solution of [4] if a.s.

$$\begin{aligned} X_t &= Z + \int_0^t a(t, X_t) dW_t + \int_0^t b(t, X_t) dt \\ \forall t &\in [0, T] \end{aligned} \quad [5]$$

provided that the right-hand side member makes sense. In particular, such a solution is continuous. The function a (resp. b) is called the diffusion (drift) coefficient of the SDE. a and b may sometimes be allowed to be random; however, this dependence has to be progressively measurable. Clearly, we can define the notion of solution $(X_t)_{t \geq 0}$ on the whole positive real axis.

We remark that those equations are called Itô SDEs. A solution of previous equation is named diffusion process.

The Lipschitz Case

The most natural framework for studying the existence and uniqueness for SDEs appears when the coefficients are Lipschitz.

A function $\gamma: [0, T] \times \mathbb{R}^m \rightarrow \mathbb{R}^d$ is said to have “polynomial growth” (with respect to x uniformly in t), if for some n there is a constant $C > 0$ with

$$\sup_{t \in [0, T]} \|\gamma(t, x)\| \leq C(1 + \|x\|^n) \quad [6]$$

The same function is said to have “linear growth” if [6] holds with $n = 1$. A function $\gamma: \mathbb{R}_+ \times \mathbb{R}^m \rightarrow \mathbb{R}^d$ is said to be “locally Lipschitz” (with respect to x uniformly in t), if for every $t \in [0, T], K > 0$, $\gamma|_{[0, T] \times [-K, K]}$ is Lipschitz (with respect to x uniformly with respect to t).

Let $a: \mathbb{R}_+ \times \mathbb{R}^{d \times m} \rightarrow \mathbb{R}^d, b: \mathbb{R}_+ \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, be Borel functions, ξ an \mathbb{R}^d -valued r.v. \mathcal{F}_0 -measurable and $(W_t)_{t \geq 0}$ be a m -dimensional (\mathcal{F}_t) -Brownian motion.

Classical fixed-point theorems allow to establish the following classical result.

Theorem 1 We suppose a and b locally Lipschitz with linear growth. Let ξ be a square-integrable r.v. that is \mathcal{F}_0 -measurable. Then [4] has a unique solution X . Moreover,

$$E\left(\sup_{t \leq T} |X_t|^2\right) < \infty$$

Remark 1

- (i) Equation [4] can be settled similarly by putting initial condition x at some time s . In that case the problem is again well stated. If $\xi \equiv x$ is a deterministic point of \mathbb{R}^d , then we will often denote by $X^{s,x}$ the solution of that problem.
- (ii) If the coefficients are only locally Lipschitz, the equation may be solved until a stopping time. If $d = 1$, it is possible to state necessary and sufficient conditions for nonexplosion (Feller test).
- (iii) The theorem above admits several generalizations. For instance, the Brownian motion can be replaced by general semimartingales, (possibly with jumps as Lévy processes).

An important role of diffusion processes is the fact that they provide probabilistic representation to PDEs of parabolic (and even elliptic) type. We will only mention here the parabolic framework.

We denote $A(t, x) = a(t, x)a(t, x)^*$, where $*$ means transposition for matrices. $(t, x) \rightarrow A(t, x) = (A_{ij}(t, x))$ is a $d \times d$ matrix-valued function. Let us consider also continuous functions $k: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d, g: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ with polynomial growth or non-negative.

Given a solution of [4], we can associate its generator $(L_t, t \in [0, T])$ setting

$$L_t f(x) = \frac{1}{2} \sum_{i,j=1}^d A_{ij}(t, x) \partial_{ij}^2 f(x) + b(t, x) \cdot \nabla f(x)$$

Feynman–Kac theorem is stated below and it provides probabilistic representation of an associated parabolic linear PDEs.

Theorem 2 Suppose there is a function $v: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ continuous with polynomial growth of class $C^{1,2}([0, T] \times \mathbb{R}^d)$ satisfying the following Cauchy problem:

$$\begin{aligned} (\partial_t v + L_t)v - kv &= g \\ v(T, x) &= f(x) \end{aligned} \quad [7]$$

Then

$$\begin{aligned} v(s, x) &= E\left(f(X_T) \exp\left(-\int_s^T k(\theta, X_\theta) d\theta\right) \right. \\ &\quad \left. - \int_s^T g(t, X_t) \exp\left\{-\int_s^t k(\theta, X_\theta) d\theta\right\} dt\right) \end{aligned}$$

for $(s, x) \in [0, T] \times \mathbb{R}^d$, where $X = X^{s,x}$. In particular, such a solution is unique.

Remark 2

- (i) In order to obtain “classical solutions” of the above Cauchy problem, one needs some conditions. It is the case, for instance, when the following ellipticity condition holds on A :

$$\exists c > 0, \forall (t, x) \in [0, T] \times \mathbb{R}^n, \forall (\xi_1, \dots, \xi_n) \in \mathbb{R}^n$$

$$\sum_{i,j} A_{ij}(t, x) \xi_i \xi_j \geq c \sum_{i=1}^d |\xi_i|^2 \quad [8]$$

In the degenerate case, it is possible to deal with viscosity solutions, in the sense of P. L. Lions. This theorem establishes an important link between deterministic PDEs and SDEs.

- (ii) A natural generalization of Feynman–Kac theorem comes from the system of forward–backward SDEs in the sense of Pardoux and Peng.
- (iii) Other types of probabilistic representation do appear in stochastic control theory through the so-called verification theorems, see for instance, Fleming and Soner (1993) and Yong and Zhou (1999). In that case, the (nonlinear) Hamilton–Jacobi–Bellman deterministic equation is represented by a controlled SDE.
- (iv) Another bridge between nonlinear PDEs and diffusions can be provided in the framework of interacting particle systems with chaos propagation, see Graham *et al.* (1996) for a survey on those problems. Among the most significant nonlinear PDEs investigated probabilistically, we quote the case of porous media equations. For instance, for a positive integer m , a solution to

$$\partial_t u = \frac{1}{2} \partial_{xx}^2 (u^{2m+1}) \quad [9]$$

can be represented by a (nonlinear) diffusion of the type, see Benachour *et al.* (1996),

$$\begin{aligned} dX_t &= u^m(s, X_s) dW_t \\ u(t, \cdot) &= \text{law density of } X_t \end{aligned} \quad [10]$$

Different Notions of Solutions

Let a and b as at the beginning of the previous section. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, a filtration $(\mathcal{F}_t)_{t \geq 0}$ fulfilling the usual conditions, an $(\mathcal{F}_t)_{t \geq 0}$ -classical Brownian motion $(W_t)_{t \geq 0}$. Let ξ be an \mathcal{F}_0 -measurable r.v. In the section “Motivation and preliminaries,” we defined the notion of solution of the following equation:

$$\begin{aligned} dX_t &= b(t, X_t) dt + a(t, X_t) dW_t \\ X_0 &= \xi \end{aligned} \quad [11]$$

This equation will be denoted by $E(a, b)$ (without initial condition). However, as we will see, the general concept of solution of an SDE is more sophisticated and subtle than in the deterministic case. We distinguish several variants of existence and uniqueness.

Definition 4 (Strong existence). We will say that equation $E(a, b)$ admits strong existence if the following holds. Given any probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a filtration $(\mathcal{F}_t)_{t \geq 0}$, an $(\mathcal{F}_t)_{t \geq 0}$ -Brownian motion $(W_t)_{t \geq 0}$, an \mathcal{F}_0 -measurable and square-integrable r.v. ξ , there is a process $(X_t)_{t \geq 0}$ solution to $E(a, b)$ with $X_0 = \xi$ a.s.

Definition 5 (Pathwise uniqueness). We will say that equation $E(a, b)$ admits pathwise uniqueness if the following property is fulfilled. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, a filtration $(\mathcal{F}_t)_{t \geq 0}$, an $(\mathcal{F}_t)_{t \geq 0}$ -Brownian motion $(W_t)_{t \geq 0}$. If two processes X, \tilde{X} are two solutions such that $X_0 = \tilde{X}_0$ a.s., then X and \tilde{X} coincide.

Definition 6 (Existence in law or weak existence). Let ν be a probability law on \mathbb{R}^d . We will say that $E(a, b; \nu)$ admits weak existence if there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a filtration $(\mathcal{F}_t)_{t \geq 0}$, an $(\mathcal{F}_t)_{t \geq 0}$ -Brownian motion $(W_t)_{t \geq 0}$, and a process $(X_t)_{t \geq 0}$ solution of $E(a, b)$ with ν being the law of X_0 .

We say that $E(a, b)$ admits weak existence if $E(a, b; \nu)$ admits weak existence for every ν .

Definition 7 (Uniqueness in law). Let ν be a probability law on \mathbb{R}^d . We say that $E(a, b; \nu)$ has a unique solution in law if the following holds. We consider an arbitrary probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a filtration $(\mathcal{F}_t)_{t \geq 0}$ on it; we consider also another probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ equipped with another filtration $(\tilde{\mathcal{F}}_t)_{t \geq 0}$; we consider an $(\mathcal{F}_t)_{t \geq 0}$ -Brownian motion $(W_t)_{t \geq 0}$, and an $(\tilde{\mathcal{F}}_t)_{t \geq 0}$ -Brownian motion $(\tilde{W}_t)_{t \geq 0}$; we suppose having a process $(X_t)_{t \geq 0}$ (resp. a process $(\tilde{X}_t)_{t \geq 0}$) solution of $E(a, b)$ on the first (resp. on the second) probability space such that both the law of X_0 and \tilde{X}_0 are identical to ν . Then X and \tilde{X} must have the same law as r.v. with values in $E = C(\mathbb{R}_+)$ (or $C[0, T]$).

We say that $E(a, b)$ has a unique solution in law if $E(a, b; \nu)$ has a unique solution in law for every ν .

There are important theorems which establish bridges among the preceding notions. One of the most celebrated is the following.

Proposition 1 (Yamada–Watanabe). Consider the equation $E(a, b)$.

- (i) Pathwise uniqueness implies uniqueness in law.
(ii) Weak existence and pathwise uniqueness imply strong existence.

A version can be stated for $E(a, b; \nu)$ where ν is a fixed probability law.

Remark 3

- (i) If a and b are locally Lipschitz with linear growth, Theorem [1] implies that $E(a, b)$ admits strong existence and pathwise uniqueness.
- (ii) If a and b are only locally Lipschitz, then pathwise uniqueness is fulfilled.

Existence and Uniqueness in Law

A way to create weak solutions of $E(1, b)$ when $(t, x) \rightarrow b(t, x)$ is Borel with linear growth is the Girsanov theorem. Suppose $d = 1$ for simplicity. Let us consider an (\mathcal{F}_t) -classical Brownian motion (X_t) . We set

$$W_t = X_t - \int_0^t b(s, X_s) ds$$

Under some suitable probability \mathbb{Q} , (W_t) is an (\mathcal{F}_t) -classical Brownian motion. Therefore, (X_t) provides a solution to $E(1, b; \delta_0)$.

We continue with an example where $E(a, b)$ does not admit pathwise uniqueness, even though it admits uniqueness in law.

Example 1 We consider the stochastic equation

$$X_t = \int_0^t \text{sign}(X_s) dW_s \quad [12]$$

with

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}$$

It corresponds to $E(a, b; \delta_0)$ with $b = 0$ and $a(x) = \text{sign}(x)$.

If $(W_t)_{t \geq 0}$ is an (\mathcal{F}_t) -classical Brownian motion, then $(X_t)_{t \geq 0}$ is $(\mathcal{F}_t)_{t \geq 0}$ -continuous local martingale vanishing at zero such that $[X]_t \equiv t$. According to Lévy characterization theorem stated earlier, X is an $(\mathcal{F}_t)_{t \geq 0}$ -classical Brownian motion. This shows in particular that $E(a, b; \delta_0)$ admits uniqueness in law. In the sequel, we will show that $E(a, b; \delta_0)$ also admits weak existence.

Let now $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, an $(\mathcal{F}_t)_{t \geq 0}$ -classical Brownian motion with respect to a filtration and $(X_t)_{t \geq 0}$ such that [12] is verified. Then $\tilde{X}_t = -X_t$ can also be shown to be a solution. Therefore, $E(a, b; \delta_0)$ does not admit pathwise uniqueness.

We continue stating a result true in the multi-dimensional case.

Proposition 3 (Stroock–Varadhan). Let ν be a probability on \mathbb{R}^d such that

$$\int_{\mathbb{R}} \|x\|^{2m} \nu(dx) < +\infty \quad [13]$$

for a certain $m > 1$. We suppose that a, b are continuous with linear growth. Then $E(a, b; \nu)$ admits weak existence.

From now on, a function $\gamma: [0, T] \times \mathbb{R}^m \rightarrow \mathbb{R}^d$ will be said Hölder-continuous if it is Hölder-continuous in the space variable $x \in \mathbb{R}^m$ uniformly with respect to the time variable $t \in [0, T]$.

Stroock and Varadhan (1979) also provide the following result, which is an easy consequence of their theorem 7.2.1.

Proposition 4 We suppose a, b both Hölder-continuous, bounded such that condition; [8] is fulfilled. Then SDE $E(a, b; \nu)$ admits weak uniqueness.

Remark 4

- (i) The Hölder condition and [8] in Proposition 4 may be relaxed and replaced with the solvability of a Cauchy problem of a parabolic PDE with suitable terminal value.
- (ii) In the case $d = 1$, if a, b are bounded and just Borel with [8] for x on each compact, then $E(a, b; \nu)$ admits weak existence and uniqueness in law. See Stroock and Varadhan (1979, exercises 7.3.2 and 7.3.3).
- (iii) If $d = 2$, the same holds as at previous point provided that moreover a does not depend on time.

We proceed with some more specifically unidimensional material stating some results from K J Engelbert and W Schmidt, who furnished necessary and sufficient conditions for weak existence and uniqueness in law of SDEs.

For a Borel function $\sigma: \mathbb{R} \rightarrow \mathbb{R}$, we first define

$$Z(\sigma) = \{x \in \mathbb{R} | \sigma(x) = 0\}$$

then we define the set $I(\sigma)$ as the set of real numbers x such that

$$\int_{x-\varepsilon}^{x+\varepsilon} \frac{dy}{\sigma^2(y)} = \infty, \quad \forall \varepsilon > 0$$

Proposition 5 (Engelbert–Schmidt criterion). Suppose that $a: \mathbb{R} \rightarrow \mathbb{R}$, that is, does not depend on time and we consider the equation without drift $E(a, 0)$.

- (i) $E(a, 0)$ admits weak existence (without explosion) if and only if

$$I(a) \subset Z(a) \quad [14]$$

- (ii) $E(a, 0)$ admits weak existence and uniqueness in law if and only if

$$I(a) = Z(a) \quad [15]$$

Remark 5

- (i) If a is continuous then, [14] is always verified. Indeed, if $a(x) \neq 0$, there is $\varepsilon > 0$ such that

$$|a(y)| > 0, \quad \forall y \in [x - \varepsilon, x + \varepsilon]$$

Therefore, x cannot belong to $I(a)$.

- (ii) Equation [14] is verified also for some discontinuous functions as, for instance, $a(x) = \text{sign}(x)$. This confirms what was affirmed previously, that is, the weak existence (and uniqueness in law) for $E(a, 0)$.
- (iii) If $a(x) = 1_{[0]}(x)$, [14] is not verified.
- (iv) If $a(x) = |x|^\alpha$, $\alpha \geq 1/2$, then

$$Z(a) = I(a) = \{0\}$$

So there is at most one solution in law for $E(a, 0)$.

- (v) The proof is technical and makes use of Lévy characterisation theorem of Brownian motion.

Results on Pathwise Uniqueness

Proposition 6 (Yamada–Watanabe). *Let $a, b: \mathbb{R}_+ \times \mathbb{R} \rightarrow \mathbb{R}$ and consider again $E(a, b)$. Suppose b globally Lipschitz and $h: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ strictly increasing continuous such that*

- (i) $h(0) = 0$;
(ii) $\int_0^\varepsilon (1/h^2)(y) dy = \infty, \quad \forall \varepsilon > 0$; and
(iii) $|a(t, x) - a(t, y)| \leq h(x - y)$.

Then pathwise uniqueness is verified.

Remark 6

- (i) In Proposition 6, one typical choice is $h(u) = u^\alpha$, $\alpha > 1/2$.
- (ii) Pathwise uniqueness for $E(a, b)$ holds therefore if b is globally Lipschitz and a is Hölder-continuous with parameter equal to $1/2$.

Corollary 1 *Suppose that the assumptions of Proposition 6 are verified and a, b continuous with linear growth. Then $E(a, b; \nu)$ admits strong existence and pathwise uniqueness, whenever ν verifies condition [13].*

Proof It follows from Propositions 6 and 3 together with Proposition 1 (ii). \square

Remark 7 Suppose $d = 1$. Pathwise uniqueness for $E(a, b)$ also holds under the following assumptions.

- (i) a, b are bounded, a is time independent and $a \geq \text{const.} > 0$, b as in Proposition 6. This result has an analogous form in the case of spacetime white noise driven SPDEs of parabolic type, as proved by Bally, Gyongy, and Pardoux in 1994.
- (ii) a independent on time, b bounded and $a \geq \text{const.} > 0$; moreover, $|a(x) - a(y)|^2 \leq |f(y) - f(x)|$ and f is increasing and bounded.

For illustration we provide some significant examples.

Example 2

$$X_t = \int_0^t |X_s|^\alpha dW_s, \quad t \geq 0 \quad [16]$$

We set $a(x) = |x|^\alpha$, $0 < \alpha < 1$. This is equation $E(a, 0)$ with $a(x) = |x|^\alpha$. According to Engelbert–Schmidt notations, we have $Z(a) = \{0\}$. Moreover

- (i) If $\alpha \geq 1/2$, then $I(a) = \{0\}$.
(ii) If $\alpha < 1/2$ then $I(a) = \emptyset$.

Therefore, according to Proposition 5, $E(a, 0)$ admits weak existence. On the other hand, if $\alpha \geq 1/2$,

$$|x^\alpha - y^\alpha| \leq h(|x - y|) \quad [17]$$

where $h(z) = z^\alpha$. According to Proposition 6, [16] admits pathwise uniqueness and by Corollary [1], also strong existence. The unique solution is $X \equiv 0$.

If $\alpha < 1/2$, $X \equiv 0$ is always a solution. This is not the only one; even uniqueness in law is not true.

Example 3 Let $a(x) = \sqrt{|x|}$, b Lipschitz. Then $E(a, b)$ admits strong existence and pathwise uniqueness. In fact, a is Hölder-continuous with parameter $1/2$ and the second item of Remark 6 applies; so pathwise uniqueness holds. Strong existence is a consequence of Propositions 3 and 1 (ii).

An interesting particular case is provided by the following equation. Let $x_0, \sigma, \delta \geq 0, k \in \mathbb{R}$. The following equation admits strong existence and pathwise uniqueness.

$$Z_t = x_0 + \sigma \int_0^t \sqrt{|Z_s|} dW_s + \int_0^t (\delta - kX_s) ds \quad [18]$$

$$t \in [0, T]$$

Equation [18] is widely used in mathematical finance and it constitutes the model of Cox–Ingersoll–Ross: the solution of the mentioned equation represents the short interest rate.

Consider now the particular case where $k = 0$, $\sigma = 2$. According to some comparison theorem for SDEs, the solution Z is always non-negative and

therefore the absolute value may be omitted. The equation becomes

$$Z_t = x_0 + 2 \int_0^t \sqrt{Z_s} dW_s + \delta t \tag{19}$$

Definition 8 The unique solution Z to

$$Z_t = x_0 + 2 \int_0^t \sqrt{Z_s} dW_s + \delta t \tag{20}$$

is called “square δ -dimensional Bessel process” starting at x_0 ; it is denoted by $BESQ^\delta(x_0)$; for fine properties of this process, see Revuz and Yor (1999, ch. IX.3).

Since $Z \geq 0$, we call δ -dimensional Bessel process starting from x_0 the process $X = \sqrt{Z}$. It is denoted by $BES^\delta(x_0)$.

Remark 8 Let $d \geq 1$. Let $W = (W^1, \dots, W^d)$ be a classical d -dimensional Brownian motion. We set $X_t = \|W_t\|$. $(X_t)_{t \geq 0}$ is a d -dimensional Bessel process.

Remark 9 If $\delta > 1$, it is possible to see that

$$X_t = W_t + \frac{\delta - 1}{2} \int_0^t \frac{ds}{X_s}$$

The Case with Distributional Drift

Pioneering work about diffusions with generalized drift was presented by N I Portenko, but in the framework of semimartingale processes. Recently, some work was done characterizing solutions in the class of the so-called Dirichlet processes, with some motivations in random irregular environment.

A useful transformation in the theory of SDE is the so-called “Zvonkin transformation.” Let (W_t) be an (\mathcal{F}_t) -classical Brownian motion. Let a (resp. b) : $\mathbb{R} \rightarrow \mathbb{R}$ (resp. C^1) be locally bounded. We suppose moreover $a > 0$. We fix $x_0 \in \mathbb{R}$. Let $(X_t)_{t \geq 0}$ be a solution of

$$\begin{aligned} X_t = x_0 + \int_0^t b(X_s) ds \\ + \int_0^t a(X_s) dW_s \end{aligned} \tag{21}$$

We set

$$\Sigma(x) = \int_0^x \frac{2b}{a^2}(y) dy$$

and we define $h : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$h(0) = 0, \quad h' = e^{-\Sigma}$$

h is strictly increasing. We set $\tilde{a}(x) = (ah')(h^{-1}(x))$, where h^{-1} is the inverse of h . We set $Y_t = h(X_t)$.

Without entering into details, the classical Itô formula allows to show that (Y_t) defines a solution of

$$\begin{aligned} dY_t &= \tilde{a}(Y_t) dW_t \\ Y_0 &= h(x_0) \end{aligned} \tag{22}$$

Now, eqn [22] fulfills the requirements of the Engelbert–Schmidt criterion so that it admits weak existence and uniqueness in law. Consequently, unless explosion, one can easily establish the same well-posedness for [21].

Zvonkin transformation also allows to prove strong existence and pathwise uniqueness results for [21]; for instance, when

- a has linear growth, and
- $y \rightarrow \int_0^y \frac{b(s)}{a^2(s)} ds$

is a bounded function.

In fact, problem [22] satisfies pathwise uniqueness and strong existence since the coefficients are Lipschitz with linear growth. Therefore, one can deduce the same for [21].

Veretennikov generalized Zvonkin transformation to the d -dimensional case in some cases which include the case $a = 1$ and b bounded Borel.

Zvonkin’s procedure suggests also to consider a formal equation of the type

$$dX_t = dW_t + \gamma'(X_t) dt \tag{23}$$

where γ is only a continuous function and so $b = \gamma'$ is a Schwartz distribution; γ could be, for instance, the realization of an independent Brownian motion of W . Therefore, eqn [23] is motivated by the study of irregular random media. When $\sigma = 1, b = \gamma',$ SDE [22], $b' = e^{-2\gamma}$ still makes sense.

Using the Engelbert–Schmidt criterion, one can see that problem [22] still admits weak existence and uniqueness in the sense of distribution laws. If Y is a solution of [22], $X = h^{-1}(Y)$ provides a natural candidate solution for [21]. R F Bass, Z-Q Chen and F Flandoli, F Russo, and J Wolf investigated generalized SDEs as [23]: in particular, they made previous reasoning rigorous, respectively, in the case of strong and weak solutions, see Flandoli *et al.* (2003).

Connected Topics

We aim here at giving some basic references about topics which are closely connected to SDEs.

Stochastic Partial Differential Equations (SPDEs)

If a SDE is a random perturbation of an ordinary differential equation, an SPDE is a random

perturbation of a PDE. Several studies were performed in the parabolic (evolution equation) and hyperbolic case (wave equation). Most of the work was done in the case of a fixed underlying probability spaces. We only quote two basic monographies which should be consulted at first before getting into the subject: the one of Walsh (1986) and the one of Da Prato and Zabczyk (1992).

However, it was possible to establish some results about weak existence and uniqueness in law for SPDEs. One possible tool was a generalization of Girsanov theorem to the case of Gaussian spacetime white noise. Weak existence for the stochastic quantization equation was proved with the help of infinite-dimensional Dirichlet forms by S Albeverio and M Röckner.

We also indicate a beautiful recent monography by Da Prato (2004) which pays particular attention to Kolmogorov equations with infinitely many variables.

Numerical Approximations

Relevant work was done in numerical approximation of solutions to SDEs and related approximations of solutions to linear parabolic equations via Feynman–Kac probabilistic representation, see Theorem 2). It seems that the stochastic simulations (of improved Monte Carlo type and related topics) for solving deterministic problems are efficient when the space dimension is greater than 4.

Malliavin Calculus

Malliavin calculus is a wide topic (see Malliavin Calculus). Relevant applications of it concern stochastic (ordinary and partial) differential equations. We only quote a monography of Nualart (1995) on those applications. Two main objects were studied.

- Given a solution of an SDE, (X_t) , sufficient conditions so that $X_t, t > 0$, has a (smooth) density $p(t, \cdot)$. Small-time asymptotics of this density, when $t \rightarrow 0$, and small-drift perturbation were performed, refining Freidlin–Ventsell large-deviation estimates.
- Coming back to SDE [11], one can conceive to consider coefficients a, b nonadapted with respect to the underlying filtration (\mathcal{F}_t) . On the other hand, the initial condition ξ may be anticipating, that is, not \mathcal{F}_0 -measurable. In that case, the Itô integral $\int_0^t a(s, X_s) dW_s$ is not defined. A replacement tool is the so-called “Skorohod integral.”

Rough Paths Approach

A very successful and significant research field is the rough path theory. In the case of dimension $d = 1$, Doss–Sussmann method allows to transform the solution of an SDE into the solution of an ordinary (random) differential equation. In particular, that solution can be seen as depending (pathwise) continuously from the driving Brownian motion (W_t) with respect to the usual topology of $C([0, T])$. Unless exceptions, this continuity does not hold in case of general dimension $d > 1$. Rough paths theory, introduced by T Lyons, allows to recover somehow this lack of continuity and establishes a true pathwise stochastic integration.

SDEs Driven by Non-semimartingales

At the moment, there is a very intense activity towards SDEs driven by processes which are not semimartingales. In this perspective, we list SDEs driven by fractional Brownian motion with the help of rough paths theory, using fractional and Young type integrals and involving finite cubic variation processes. Among the contributors in that area we quote L Coutin, R Coviello, M Errami, M Gubinelli, Z Qian, F Russo, P Vallois, and M Zähle.

See also: Fractal Dimensions in Dynamics; Image Processing: Mathematics; Interacting Stochastic Particle Systems; Lagrangian Dispersion (Passive Scalar); Malliavin Calculus; Path Integrals in Noncommutative Geometry; Quantum Dynamical Semigroups; Quantum Fields with Indefinite Metric: Non-Trivial Models; Random Dynamical Systems; Random Walks in Random Environments; Stochastic Hydrodynamics; Stochastic Resonance.

Further Reading

- Bally V, Gyongy I, and Pardoux E (1994) White noise driven parabolic SPDEs with measurable drift. *Journal of Functional Analysis* 120: 484–510.
- Benachour S, Chassaing P, Roynette B, and Vallois P (1996) Processus associés l'équation des milieux poreux. *Annali Scuola Normale Superiore di Pisa, Classe di Scienze* 23(4): 793–832.
- Bouleau N and Lépine D (1994) *Numerical Methods for Stochastic Processes*. New York: Wiley.
- Da Prato G (2004) *Kolmogorov Equations for Stochastic PDEs*. Basel: Birkhäuser.
- Da Prato G and Zabczyk J (1992) *Stochastic Equations in Infinite Dimensions*, Encyclopedia of Mathematics and its Applications, vol. 44. Cambridge: Cambridge University Press.
- Flandoli F, Russo F, and Wolf J (2003) Some stochastic differential equations with distributional drift. Part I. General calculus. *Osaka Journal of Mathematics* 40(2): 493–542.
- Fleming WH and Soner M (1993) *Controlled Markov Processes and Viscosity Solutions*. New York: Springer.
- Graham C, Kurtz Th G, Méléard S, Protter Ph E, Pulvirenti M, and Talay D (1996) Probabilistic models for nonlinear partial

- differential equations. In: Talay and Tubaro L (eds.) *Lectures given at the 1st Session and Summer School held in Montecatini Terme*, 22–30 May 1995, Lecture notes in Mathematics, vol. 1627. Springer-Verlag, Centro Internazionale Matematico Estivo (C.I.M.E.), Florence, 1996.
- Karatzas I and Shreve SE (1991) *Brownian Motion and Stochastic Calculus*, 2nd edn. New York: Springer.
- Kloeden PE and Platen E (1992) *Numerical Solutions of Stochastic Differential Equations*. Berlin: Springer.
- Lamberton D and Lapeyre B (1997) *Introduction au Calcul Stochastique et Applications la Finance*. Paris: Collection Ellipses.
- Ma J and Yong J (1999) *Forward-Backward Stochastic Differential Equations and Their Applications*, Lecture Notes in Mathematics, vol. 1702. Berlin: Springer.
- Nualart D (1995) *The Malliavin Calculus and Related Topics*. New York: Springer.
- Øksendal B (2003) *Stochastic Differential Equations. An Introduction with Applications*. Sixth edition, Universitext. Berlin: Springer-Verlag.
- Pardoux E (1998) *Backward stochastic differential equations and viscosity solutions of systems of semilinear parabolic and elliptic PDEs of second order*. Stochastic analysis and related topics, VI (Geilo, 1996), 79–127, Progr. Probab., 42, Boston: Birkhäuser, 1998.
- Protter P (1992) *Stochastic Integration and Differential Equations. A new approach*. Berlin: Springer.
- Lyons T and Qian Z (2002) *Controlled Systems and Rough Paths*. Oxford: Oxford University Press.
- Revuz D and Yor M (1999) *Continuous Martingales and Brownian Motion*, Third edition. Berlin: Springer-Verlag.
- Stroock D and Varadhan SRS (1979) *Multidimensional Diffusion Processes*. Berlin-New York: Springer.
- Walsh JB (1986) An introduction to stochastic partial differential equations. In: *Ecole d'été de probabilités de Saint-Flour, XIV – 1984*, pp. 265–439. Lecture Notes in Mathematics. vol. 118, Springer.
- Yong J and Zhou XY (1999) *Stochastic Controls: Hamiltonian Systems and HJB Equations*. New York: Springer.

Stochastic Hydrodynamics

B Ferrario, Università di Pavia, Pavia, Italy

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Mathematical models in hydrodynamics are introduced to describe the motion of fluids. The basic equations for Newtonian incompressible fluids are the Euler and the Navier–Stokes equations, for inviscid and viscous fluids, respectively. For a given set of body forces acting on the fluid, these nonlinear partial differential equations (PDEs) model the evolution in time of the velocity and pressure at each point of the fluid, given the initial velocity and suitable boundary conditions (see Partial Differential Equations: Some Examples). The equations of hydrodynamics offer challenging mathematical problems, like proving the existence and uniqueness of solutions, determining their regularity, their asymptotic behavior for large time, and their stability. To gain some insight into the behavior of fluids, stochastic analysis is introduced into hydrodynamics. In fact, there are various attempts to describe turbulent regime (see Turbulence Theories). But, analyzing individual solutions that determine the flow at any time, for a given initial condition, is a desperate task, since the dynamics in a turbulent regime is chaotic and highly unstable. This is a particular chaotic motion with some characteristic statistical properties (see Monin and Yaglom (1987)). The aim of a statistical description of turbulent flow is to single out some

relevant collective properties of the flow that, hopefully, make it possible to grasp the salient features of the dynamics. In this sense, stochastic hydrodynamics is germane to the kinetic gas theory. In the next section we shall review a typical topic of stochastic hydrodynamics, the evolution of probability measures. Results on stationary probability measures will be given in the subsequent sections.

Another characteristic of turbulent flows is the lack of space regularity of the velocity field. We shall introduce in the section “The stochastic Navier–Stokes equations” a stochastic model of turbulence, which exhibits lack of regularity of the solutions.

The Euler equations are a singular limit of the Navier–Stokes equations, since they are first order, instead of second-order PDEs. It is little surprise if they involve different mathematical techniques. A full section will be devoted to a discussion of Euler equations and another to the Navier–Stokes equations. Statistics of an inviscid flow, when approximated by vortex motion, will be described in the final section.

Statistical Solutions

Let $u(t, x)$ be the fluid velocity at time t and point $x \in D \subseteq \mathbb{R}^d$; since the initial velocity is always affected by experimental errors, it is reasonable to assign a measure ν determining the probability that the initial velocity belongs to a Borel set Γ of the space \mathbb{H} of all admissible velocity fields $u = u(x)$.

A spatial statistical solution is a family of probability measures $\mu(t, \cdot), t \geq 0$, each supported

on the set \mathbb{H} such that, given any Borel set Γ in \mathbb{H} , we have

$$\text{Prob}\{u(t, x) \in \Gamma\} = \mu(t, \Gamma), \quad \forall t > 0 \quad [1]$$

with the initial condition $\mu(0, \Gamma) = \nu(\Gamma)$. The construction and analysis of statistical solutions $\mu(t, \cdot)$ is one of the crucial mathematical problems in stochastic hydrodynamics (see, e.g., Vishik and Fursikov (1988)).

Hopf gave the first mathematical formulation of the problem of describing turbulent flows by statistical solutions. The first result on the existence of statistical solutions is by Foias in 1973. Hopf (1952) presented an equation in variational derivatives satisfied by the characteristic functional $\chi(t, \phi)$ of the family of measures $\mu(t, \cdot)$ associated with the Navier–Stokes equations. The characteristic functional $\chi(t, \phi)$ is the Fourier transform of the measure $\mu(t, \cdot)$:

$$\chi(t, \phi) = \int_{\mathbb{H}} e^{i\langle \phi, u \rangle} \mu(t, du) \quad [2]$$

defined for any smooth test function ϕ .

We now derive the evolution equation for $\chi(t, \phi)$, by assuming that the dynamics takes place in the phase space \mathbb{H} and follows the nonlinear equation

$$\frac{du}{dt} = F(u) \quad [3]$$

If $u^\nu(t)$ is the solution started from ν at time $t=0$, then its probability distribution is represented by the time-evolved measure $\mu(t, \cdot)$. Therefore, we have that

$$\int_{\mathbb{H}} e^{i\langle \phi, u \rangle} \mu(t, du) = \int_{\mathbb{H}} e^{i\langle \phi, u^\nu(t) \rangle} \mu(0, d\nu) \quad [4]$$

Differentiating in time, we obtain

$$\begin{aligned} \frac{d}{dt} \chi(t, \phi) &= \int_{\mathbb{H}} e^{i\langle \phi, u^\nu(t) \rangle} i \langle \phi, F(u^\nu(t)) \rangle \mu(0, d\nu) \\ &= i \int_{\mathbb{H}} e^{i\langle \phi, v \rangle} \langle \phi, F(v) \rangle \mu(t, dv) \end{aligned} \quad [5]$$

The last integral is uniquely determined by χ , since the measure $\mu(t, \cdot)$ is uniquely determined by $\chi(t, \phi)$. We denote by $\Phi_\chi(t, \phi)$ the last integral in [5]. The evolution equation thus obtained for the characteristic functional χ is

$$\frac{d}{dt} \chi(t, \phi) = i \Phi_\chi(t, \phi), \quad \forall \phi \quad [6]$$

This is called the Hopf equation associated with the dynamical system [3].

Another way to analyze the evolution of measures is through the moments; instead of the measure $\mu(t, \cdot)$

describing the spatial statistical solution, we deal with the moments of $\mu(t, \cdot)$ of any order. For a nonlinear dynamics [3], the moments equations are an infinite chain of coupled equations, the so-called Friedman–Keller equations.

A prominent role among statistical solutions is played by stationary solutions. They contain all the statistical information in the case of equilibrium in time. We have that the characteristic functional of an invariant measure is constant in time. Therefore,

$$\frac{d}{dt} \chi(t, \phi) = 0$$

Bearing in mind equation [5], this is equivalent to say that the signed measure $\langle \phi, F(v) \rangle \mu(t, dv)$ vanishes, for any test function ϕ and time t . Setting $t=0$, we obtain that an invariant measure ν in the space \mathbb{H} satisfies the Liouville equation

$$\int_{\mathbb{H}} \langle \phi(v), F(v) \rangle d\nu(v) = 0 \quad [7]$$

for appropriate test functions ϕ . This equation is also called the relation of infinitesimal invariance and the measure ν is said to be infinitesimally invariant.

The stationary measures are natural candidates to describe the statistical asymptotic behavior of the system when $t \rightarrow \infty$. Notice that, in a chaotic system two motions that are arbitrarily close to one another at $t=0$ can evolve in completely different ways. So, to describe satisfactorily the dynamics we take average over a big number of experiments. This is the so-called ensemble average. These averages are assumed to be with respect to an invariant measure μ . The invariant measures must exist and either they are unique or at most one has physical meaning and enters in the functional integral defining the ensemble average. According to the ergodic principle (an assumption not yet proved in hydrodynamics), ensemble averages replace long-time averages: for every initial velocity field ν , except for a set of initial values negligible in some sense, the time average of an observable ψ tends, as time goes to infinity, to the ensemble average

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \psi(u^\nu(t)) dt = \int_{\mathbb{H}} \psi d\mu \quad [8]$$

However, it is extremely difficult to prove the existence of stationary probability measures for the Navier–Stokes equations solving directly equation [7]. The situation is formally the same as in equilibrium statistical mechanics, where the Liouville equation is in fact solved, leading to the Boltzmann–Gibbs distribution. However, the results in statistical hydrodynamics are far from being satisfactory.

Recent studies to prove the existence of invariant measures for the Navier–Stokes equations are based on stochastic models (see the section “The stochastic Navier–Stokes equations”). On the other hand, for the Euler equations it is possible to construct formally invariant measures, by means of invariant quantities of the classical motion (see the next section).

Finally, we point out that there are techniques using invariant measures to show some results for the time evolution (e.g., the motion exists for almost all initial values with respect to an invariant measure).

The Euler Equations

We start recalling some basic facts on Euler equations (see Incompressible Euler Equations: Mathematical Theory).

The motion of an inviscid, incompressible, and homogeneous fluid is described by the Euler equations, which in Eulerian coordinates read as

$$\begin{aligned} \frac{\partial u}{\partial t} + (u \cdot \nabla)u + \nabla p &= f \\ \nabla \cdot u &= 0 \quad \text{in } D \\ u \cdot n &= 0 \quad \text{on } \partial D \end{aligned} \quad [9]$$

where, at time $t \geq 0$ and position $x \in D$, $u = u(t, x)$ is the vector velocity, $p = p(t, x)$ the hydrodynamic pressure. The units have been chosen so that the mass density $\rho = 1$. ∇ denotes the nabla vector operator so

$$\begin{aligned} u \cdot \nabla &= \sum_{j=1}^d u_j \frac{\partial}{\partial x_j} \\ \nabla \cdot u &= \sum_{j=1}^d \frac{\partial u_j}{\partial x_j} \\ \nabla p &= \left(\frac{\partial p}{\partial x_1}, \dots, \frac{\partial p}{\partial x_d} \right) \end{aligned}$$

Finally, f denotes the external force. If the spatial domain D has a boundary ∂D , then the velocity is assumed to be tangent to the boundary (n denotes the exterior normal vector to the boundary). Some initial condition u_0 at time $t = 0$ is assigned.

When $f = 0$, there are invariant quantities for system [9]. In the literature, there are many works suggesting a Gaussian stationary statistics (see, e.g., the paper by Kraichnan (1980)). We consider invariants that are quadratic in the velocity so as

to construct (formally) invariant measures of Gibbs type: the *energy*

$$\mathcal{E}(u) := \frac{1}{2} \int_D |u|^2 dx$$

and, only in the two-dimensional case ($d = 2$), the *enstrophy*

$$\mathcal{S}(u) := \frac{1}{2} \int_D |\text{curl } u|^2 dx$$

(with $\text{curl } u = \nabla^\perp \cdot u \equiv \partial u_2 / \partial x_1 - \partial u_1 / \partial x_2$ for $d = 2$).

It is natural to look for velocity fields in the following function spaces: the space H^0 of finite kinetic energy and the space H^1 of finite enstrophy. Clearly, the admissible fields should also obey the boundary conditions and divergence-free condition. If P is the projection operator onto the space of divergence-free vectors, and B is the bilinear form $B(u, v) := P[(u \cdot \nabla)v]$, the Euler equations can be given the structure of an evolution,

$$\frac{du}{dt} = -B(u, u) \quad [10]$$

obtained by applying the projection operator P to the first equation in [9]. The pressure disappears and can be regarded as a Lagrange multiplier associated with the divergence-free constraint ($\nabla \cdot u = 0$); it can be fully recovered once the velocity field is known. The dynamics is considered in the phase space of divergence-free velocity vectors \mathbb{H} (a large space containing H^0 and H^1), which is an infinite-dimensional functional space. More precisely, identifying H^0 with its dual $(H^0)'$, we introduce the Gelfand's triplet

$$H^1 \subset H^0 \subset (H^1)' = H^{-1}$$

The space H^α , with $\alpha = 1, 2, \dots$, are the usual Sobolev spaces but with the additional divergence-free and boundary conditions. For $\alpha > 0$ noninteger, the spaces H^α are defined by interpolation, whereas those with $\alpha < 0$ by duality. As usual, regularity in space is related to the spaces H^α with higher exponent α . We have that $\mathbb{H} = \bigcup_{\alpha \in \mathbb{R}} H^\alpha$.

Invariance of \mathcal{E} and \mathcal{S} can be proved resorting to eqn [9] and assuming that u is a smooth vector field. For instance,

$$\begin{aligned} \frac{d}{dt} \mathcal{E}(u(t)) &= \frac{d}{dt} \frac{1}{2} \int_D |u|^2 dx = \int_D u \cdot \frac{\partial u}{\partial t} dx \\ &= - \int_D u \cdot [(u \cdot \nabla)u] dx - \int_D u \cdot \nabla p dx \end{aligned}$$

By integrating by parts and bearing in mind the divergence-free condition and the boundary condition, we conclude that

$$\frac{d}{dt} \mathcal{E}(u) = 0$$

In the same way, the invariance of \mathcal{S} can be proved.

As a consequence, the following Gibbs measures which are defined on the space H

$$\begin{aligned} \mu_{\mathcal{E}}(du) &= \frac{1}{Z_{\mathcal{E}}} e^{-\mathcal{E}(u)} du \\ \mu_{\mathcal{S}}(du) &= \frac{1}{Z_{\mathcal{S}}} e^{-\mathcal{S}(u)} du \end{aligned} \quad [11]$$

are heuristically invariant in time. In [11], $Z_{\mathcal{E}}$ and $Z_{\mathcal{S}}$ are the partition functions, that is, they are normalization constants needed to guarantee that $\mu_{\mathcal{E}}$ and $\mu_{\mathcal{S}}$ are genuine probability measures (e.g., $Z_{\mathcal{E}} = \int_H e^{-\mathcal{E}(u)} du$).

Actually, these measures μ solve the Liouville equation

$$\int_H \langle \phi(u), B(u, u) \rangle d\mu(u) = 0 \quad [12]$$

for any test function ϕ , cylindrical, infinitely differentiable, bounded, and with bounded derivatives.

On the other hand, the (global and not only infinitesimal) invariance means that if there exists a global flow in time which is well defined in a phase space of full measure μ , then the measure μ is invariant under this dynamics. The measures $\mu_{\mathcal{E}}$ and $\mu_{\mathcal{S}}$ are centered Gaussian measures whose support is in a space larger than H^0 , as can be proved by standard methods in the theory of Gaussian measures on infinite-dimensional spaces. By the very definition, $\mu_{\mathcal{E}}$ is a cylindrical measure in H^0 and $\mu_{\mathcal{S}}$ is cylindrical in H^1 . Then the support of $\mu_{\mathcal{E}}$ is any Hilbert space \tilde{H} such that $H^0 \subset \tilde{H}$ is a Hilbert-Schmidt embedding, and the support of $\mu_{\mathcal{S}}$ is any space \tilde{H} such that $H^1 \subset \tilde{H}$ is a Hilbert-Schmidt embedding. When the spatial dimension d is 2, $\text{supp}(\mu_{\mathcal{E}}) = \cap_{\alpha < -1} H^{\alpha}$ and $\text{supp}(\mu_{\mathcal{S}}) = \cap_{\alpha < 0} H^{\alpha}$. When d is 3, $\text{supp}(\mu_{\mathcal{E}}) = \cap_{\alpha < -3/2} H^{\alpha}$.

Moreover, $\mu_{\mathcal{E}}(H^0) = \mu_{\mathcal{S}}(H^0) = 0$, that is, the space of finite energy H^0 is negligible with respect to these measures. Let us show this property for the “enstrophy measure” $\mu_{\mathcal{S}}$ when $d=2$. Let $\{e_j\}_{j=1}^{\infty}$ be a complete orthonormal system in H^0 . Hence, for $u = \sum_j u_j e_j$, we have $\|u\|_{H^0}^2 = \sum_j |u_j|^2$ and $\|u\|_{H^1}^2 = \sum_j \lambda_j |u_j|^2$ (with $0 < \lambda_1 \leq \lambda_2 \leq \dots$ and $\lambda_j \sim j$ as $j \rightarrow \infty$). Keeping in mind its definition, the measure $\mu_{\mathcal{S}}$ can be considered as a measure on the space of the sequences $\{u_j\}_j$ and written as an infinite product of one-dimensional centered Gaussian measures

$$\mu_{\mathcal{S}}(du) = \otimes_j \frac{1}{\sqrt{2\pi\lambda_j^{-1}}} e^{-(\lambda_j/2)|u_j|^2} du_j \quad [13]$$

The energy is

$$\mathcal{E}(u) = \frac{1}{2} \sum_j |u_j|^2$$

and the renormalized energy is

$$:\mathcal{E}:(u) = \frac{1}{2} \sum_j \left(|u_j|^2 - \int |u_j|^2 \mu_{\mathcal{S}}(du) \right)$$

Since, as can be easily shown $\int (:\mathcal{E}:(u))^2 \mu_{\mathcal{S}}(du) < \infty$, $:\mathcal{E}:(u)$ is finite for $\mu_{\mathcal{S}}$ -almost every u . On the contrary, since $\sum_j \int |u_j|^2 \mu_{\mathcal{S}}(du) = \sum_j \lambda_j^{-1} = +\infty$, $\mathcal{E}(u)$ is infinite for $\mu_{\mathcal{S}}$ -almost every u .

We also note in passing that, for any $\gamma > 0$ and $\beta > -\gamma$

$$\int_H e^{-\beta:\mathcal{E}:(u)} \frac{e^{-\gamma\mathcal{S}(u)}}{Z} du < \infty$$

so that

$$\mu_{\mathcal{S}}^{(\beta),(\gamma)}(du) = \frac{e^{-\beta:\mathcal{E}:(u) - \gamma\mathcal{S}(u)}}{\int e^{-\beta:\mathcal{E}:(u) - \gamma\mathcal{S}(u)} du} du \quad [14]$$

is a probability measure, which is infinitesimally invariant for the Euler flow.

Since the space of finite-energy velocity is negligible with respect to these measures, it is necessary to replace the classical solutions having finite energy with generalized solutions. This is not an easy task in the three-dimensional case, whereas some results have been proved for the two-dimensional problem, where the following existence result holds. Let us analyze the quadratic term $B(u, u) = -P[(u \cdot \nabla)u] \cdot (u \cdot \nabla)u$ can be rewritten as $\nabla(u \otimes u)$, taking in account the divergence-free condition. Trivially, we have that $\nabla(u \otimes u) = \nabla(u \otimes u - :u \otimes u:)$, where $:u \otimes u: = \int u \otimes u; \mu_{\mathcal{S}}(du)$. We consider the quadratic expression $(u \otimes u - :u \otimes u:)$. This is integrable with respect to the measure $\mu_{\mathcal{S}}$ in the sense that

$$\int \|u \otimes u - :u \otimes u:\|_{H^{-\varepsilon}}^2 \mu_{\mathcal{S}}(du) < \infty \quad [15]$$

for any $\varepsilon > 0$. We remark that this property is similar to the integrability of the renormalized energy, which is a quadratic expression as well. This implies that the $H^{-1-\varepsilon}$ -norm of $\nabla(u \otimes u)$ is integrable with respect to the measure $\mu_{\mathcal{S}}$. Therefore, $B(u, u)$ is defined for $\mu_{\mathcal{S}}$ -a.e. u .

Now, let us replace eqn [10] with a system of infinite equations for all the components u_j with respect to the orthonormal basis $\{e_j\}_j$, obtained by taking the scalar product with e_j of both sides of eqn [10]:

$$\frac{du_j}{dt} = B_j(u, u), \quad j = 1, 2, \dots \quad [16]$$

Each component $B_j(u, u)$ is defined for μ_S -a.e. u . These estimates lead to define a weak solution (see Albeverio and Cruzeiro (1990)):

Theorem 1 *Let $d=2$. There exists a flow $U(t, \omega)$ defined on a probability space (Ω, \mathcal{F}, P) with values in $H^{-\varepsilon-1}$ for any $\varepsilon > 0$, $U(\cdot, \omega) \in C(\mathbb{R}, H^{-\varepsilon-1})$ P -a.e. ω , such that for each component U_j we have*

$$\begin{aligned} U_j(t, \omega) &= U_j(0, \omega) + \int_0^t B_j(U(s, \omega), U(s, \omega)) ds, \\ P - a.e. \omega. \quad \forall t \in \mathbb{R} \end{aligned}$$

Moreover, the measure μ_S is invariant under this flow.

We point out that uniqueness is an open problem also for $d=2$. But already in the classical analysis of the Euler equations in a bounded domain, uniqueness for initial velocity of finite energy is not known. Working with the measure μ_ε is even worse, especially when $d=3$, because its support is a larger space within which more irregular velocity vectors live. The more irregular the spaces where the flow lives, the more difficult is to handle the nonlinear term $B(u, u)$.

On the other hand, for $d=1$, the mathematical analysis is much easier. For instance, it can be proved (see Robert (2003)) that the one-dimensional inviscid Burgers equation on the line

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left(\frac{1}{2} u^2 \right) = 0 \quad [17]$$

has intrinsic invariant statistical solution, given by a class of Lévy's processes with negative jumps.

The Stochastic Navier–Stokes Equations

The Navier–Stokes equations describe advection with velocity u and diffusion with kinematic viscosity $\nu > 0$ (see Viscous Incompressible Fluids: Mathematical Theory)

$$\begin{aligned} \frac{\partial u}{\partial t} - \nu \Delta u + (u \cdot \nabla) u + \nabla p &= f \\ \nabla \cdot u &= 0 \\ u &= 0 \quad \text{on } \partial D \end{aligned} \quad \text{in } D \quad [18]$$

where Δ is the Laplace operator. Nonslip boundary conditions are assumed. Although the Euler equations [9] are formally obtained from [18] by setting $\nu=0$, the presence of the second-order operator $-\nu\Delta$ makes the analysis needed to prove the existence, uniqueness, and regularity of solutions

easier than for the Euler equations. However, at variance with the Euler equations, the Navier–Stokes equations do not possess invariants, since the viscosity dissipates energy. Hence, it is difficult to find explicit expressions of invariant measures for the deterministic Navier–Stokes equations, except the trivial invariant measures concentrated on a stationary solution. However, as soon as a stochastic force is introduced in these equations, it is possible to have nontrivial invariant measures. It is impossible to review here the wide literature concerning the stochastic Navier–Stokes equations and we confine ourselves to make some remarks. Most results are concerned with proving the existence and/or uniqueness of an invariant measure μ , without giving an explicit representation, apart some attempts like Gallavotti (2002), where a formal representation of stationary distributions is given in terms of functional integrals. Some properties of the not explicit invariant measures are given like, for instance, estimates of moments, exponential convergence of the statistical solution for large time.

Stochastic forces can enter in the Navier–Stokes equations in different ways. We can consider randomness in the forcing term, so that the force f in [18] has a deterministic component which represents its mean varying slowly and a stochastic one, which accounts small fluctuations around the mean and varying very rapidly. Alternatively, since the molecules are not rigidly connected to one another in the fluid, they are subjected to fluctuations. A complete description of fluctuations relating the microscopic and macroscopic motion is not achieved at present. However, we shall introduce some models for which rigorous mathematical results can be proved.

The first part of this section concerns the Navier–Stokes equations with noise n :

$$\begin{aligned} \frac{\partial u}{\partial t} - \nu \Delta u + (u \cdot \nabla) u + \nabla p &= n \\ \nabla \cdot u &= 0 \end{aligned} \quad [19]$$

for which invariant measures exist, one of which can be ergodic provided that the noise is suitably chosen. In the second part, a Navier–Stokes-type stochastic system is described, which has irregular solutions, as expected in turbulence.

Let us introduce the stochastic Navier–Stokes equations with time white noise. The first equation in [19] is an Itô equation:

$$\partial_t u + [-\nu \Delta u + (u \cdot \nabla) u + \nabla p] = \partial_t w \quad [20]$$

Here $w = w_{(1)}, \dots, w_{(d)}$ is a Brownian motion, that is, its time derivative $n = \partial w / \partial t$ is a Gaussian

stochastic field with zero mean and correlation function given by

$$\begin{aligned} \mathbb{E}[n_{(j)}(t, x) n_{(k)}(t', x')] \\ = \delta_{jk} q(x - x') \delta(t - t') \end{aligned} \quad [21]$$

for $j, k = 1, \dots, d$.

We shall use the differential form for the Itô equation [20] always understood in the integral form

$$\begin{aligned} u(t) - u(0) + \int_0^t [-\nu \Delta u(s) + (u(s) \cdot \nabla) u(s) \\ + \nabla p(s)] ds = w(t) \end{aligned} \quad [22]$$

Modeling perturbations by a white noise process represents the first step to understand how a random perturbation acts in the mathematical equations, rather than a good physical or numerical model. The first results are in a paper by Bensoussan and Temam (1973).

Obviously, the regularity of the solutions depends on the spatial covariance q of the noise.

Let us consider the following cases.

- $q = \delta$: the noise is white also in space.

An invariant measure is known explicitly. Indeed, assume periodic boundary conditions on the square ($d=2$) or the cube ($d=3$) D , which makes the spatial domain a torus. In this case, the Euler and Navier–Stokes equations are set in the same functional spaces. The generator of the stochastic Navier–Stokes equations [20] corresponds to the sum of the generator of the Euler equations [9] and of the stochastic Stokes equations

$$\begin{aligned} \partial_t u &= [\nu \Delta u - \nabla p] + \partial_t w \\ \nabla \cdot u &= 0 \end{aligned} \quad [23]$$

Since the first equation in [23] is linear in the unknown velocity u , the Stokes system has a unique invariant measure which is a centered Gaussian measure. In particular, when the noise is a space-time white noise and $d=2$, this is the invariant measure [14] of the enstrophy:

$$\mu_S^{(0), (2\nu)}(du) = \frac{1}{Z} e^{-2\nu S(u)} du$$

On a bidimensional torus, it is proved that this measure is not only infinitesimally invariant, but also globally invariant for a unique flow [20] defined for $\mu_S^{(0), (2\nu)}$ -a.e. initial velocity. We recall that initial velocities of finite energy are negligible with respect to the measure $\mu_S^{(0), (2\nu)}$.

- q more regular than above, that is, the noise is colored in space.

As soon as the forcing term is more regular in space, the Navier–Stokes system has a solution of finite energy. These are solutions close to those of the deterministic equation. Techniques similar to those used to prove the existence and/or uniqueness of solutions for the deterministic equations work also in the stochastic case with an additive noise (or even a multiplicative noise) to get weak or strong solutions. Global existence in the space H^0 is proved for $d=2, 3$ and uniqueness only for $d=2$, as is the case for the deterministic Navier–Stokes equations.

The interesting feature is that by adding a noise which acts on all the components with respect to a Hilbert basis (or at least on many components), the stochastic Navier–Stokes system has a unique invariant measure, which is ergodic. This is proved for the spatial dimension $d=2$. By means of the Krylov–Bogoliubov’s method, existence of at least an invariant measure is proved by compactness of a family of averaged measures; the limit measures are stationary measures. But, when many modes are perturbed by a noise, there is a mixing effect on the dynamics, avoiding existence of many stationary measures. For the spatial dimension $d=2$, the best result in this context is in Hairer and Mattingly (2004), where the noise acts on very few modes. For the spatial dimension $d=3$, the result in Da Prato and Debussche (2003) shows the existence of an invariant measure; even if there is no uniqueness of the solutions (as in the deterministic case), by a selection principle, they construct a transition semigroup, which has a unique invariant measure, ergodic and strongly mixing.

Mathematical proofs are given for very different noises. (The reader is urged to consult, among the others, the papers by E, Mattingly and Sinai; Flandoli and Maslowski; Mikulevicius and Rozovskii; Vishik and Fursikov. The latter authors study also statistical solutions in two and three dimensions. For a kick noise $n = \sum_k \delta(t - k) q_k(x)$ in equations [19], there are results for $d=2$ by Bricmont, Kupiainen and Lefevere; Kuksin and Shirikyan.)

We conclude that, as far as invariant measures and their ergodicity are concerned, the stochastic Navier–Stokes equations have richer results than the deterministic Navier–Stokes equations. It is appealing to investigate the limit as the intensity of the noise goes to zero, so as to recover the deterministic equation. Now, think of equation [19] with a noise εn , for n fixed and $\varepsilon \rightarrow 0$. Due to the sensitive dependence on initial conditions, even a small noise may have important effects on the dynamics. A conjecture by Kolmogorov is that the unique invariant measure μ_ε tends, when $\varepsilon \rightarrow 0$, to a specific measure, the so-called Kolmogorov measure, which

would enter into the ergodic principle. This is a difficult problem, not yet solved.

We also mention the analysis of the inviscid limit. Kuksin (2004) showed that the solution u_ν of the two-dimensional stochastic Navier–Stokes equations

$$\frac{\partial u}{\partial t} - \nu \Delta u + (u \cdot \nabla)u + \nabla p = \sqrt{\nu}n, \quad 0 < \nu \leq 1 \quad [24]$$

on the torus converges in distribution to a stationary solution of the Euler equations. Here n is a random force white in time and smooth in space. More precisely, for each subsequence u_{ν_j} ,

$$\lim_{\nu_j \rightarrow 0} \lim_{T \rightarrow \infty} u_{\nu_j}(T+t) = U(t) \quad [25]$$

and almost every trajectory of the nontrivial limit process U solves the Euler equations [9] without the forcing term. Moreover, the process U keeps memory of some features of the noise force n , since the mean values of the enstrophy and of the energy of U depend on the noise n .

We now present the second part on stochastic models for viscous fluids. In his 1884 paper, Reynolds introduced the decomposition of turbulent flow into mean and fluctuating flows. The equations obtained are difficult to study. We shall show now a tractable model for a one-dimensional problem ($d=1$) with a suitable model of fluctuations. Decompose the velocity field into the sum of a mean flow \bar{u} and a fluctuation δ

$$u = \bar{u} + \delta$$

The fluctuation is assumed to be highly irregular; it is reasonable to model it by a stochastic process. If we choose

$$\delta = b \frac{dw}{dt}$$

where b is a given velocity field and dw/dt is white noise, then the motion of the fluid is governed by a stochastic equation of Itô type. Indeed, the Navier–Stokes equations are balance equations of linear momentum:

$$\frac{Du}{Dt} = \nu \Delta u - \nabla p \quad [26]$$

where Du/Dt is the material time derivative along the trajectory of a particle which is at time t in position $x(t)$ moving with velocity u (so $u(x(t)) = (dx/dt)(t)$):

$$\frac{Du}{Dt} = \frac{d}{dt} u(t, x(t)) = \frac{\partial u}{\partial t} + (u \cdot \nabla)u \quad [27]$$

According to the mathematical model for the fluctuation, we have

$$dx(t) = \bar{u}(t, x(t))dt + b(x(t))dw(t) \quad [28]$$

Therefore, $D\bar{u}$ is computed by means of Itô's formula

$$\begin{aligned} D\bar{u}(t, x(t)) &= \frac{\partial \bar{u}}{\partial t} dt + \sum_{k=1}^d \frac{\partial \bar{u}}{\partial x_k} dx_k(t) \\ &\quad + \frac{1}{2} \sum_{k,s=1}^d \frac{\partial^2 \bar{u}}{\partial x_k \partial x_s} b_k b_s dt \end{aligned} \quad [29]$$

This leads to the stochastic Navier–Stokes-type equations (we neglect the overline symbol)

$$\begin{aligned} d_t u + [-\nu \Delta u + (u \cdot \nabla)u + \nabla p + \frac{1}{2} Qu] dt \\ = -(b \cdot \nabla)u dw(t) \\ \nabla \cdot u = 0 \end{aligned} \quad [30]$$

where Q is the second-order differential operator given by the last term in [29].

Rigorous mathematical results for the above equations have been proved for the one-dimensional case, that is, the Burgers equations on the line. Given an initial velocity of finite energy $u_0 \in H^0$, there exists a unique solution $u \in C([0, T]; H^0) \cap L^2(0, T; H^1)$ (P-a.s.). But it can be shown that for a more regular initial velocity there is no higher regularity of the solution of eqn [30], if $b \neq 0$. This means that these stochastic Burgers equations cannot have too regular nontrivial solutions, as expected in turbulent motion.

Statistics of Vortices and Bidimensional Turbulence

Onsager (1949) proposed to investigate bidimensional turbulent flows, extending in a rigorous way to hydrodynamics the statistical mechanics approach of Boltzmann. If we are interested in flows of finite energy, the results of the section “The Euler equations” provide no answer to the problem. Another way to proceed is by approximating the Euler equations in a suitable way. Actually, in a two-dimensional turbulent flow, there appears a large-scale organization leading to coherent structures. These are hydrodynamical vortices, whose dynamics is governed by the Euler equations. Onsager suggested to approximate the continuous Euler equations by a great (but finite) number of point vortices. This leads to a finite-dimensional Hamiltonian system, to which the methods of statistical mechanics can be successfully applied. Of course, the crucial point is to pass to the limit, to

recover the continuous system. But there are many different ways to approximate a continuous vorticity by a cloud of point vortices and different approximations may lead to very different statistical equilibrium states.

We present here the approach presented in Lions (1997). To get an idea of a completely different approximation, see, for example, Robert (2003).

Let D be a bounded open smooth simply connected subset of \mathbb{R}^2 . Then there exists a function ψ (the stream function) such that $u = \nabla^\perp \psi$ and $\psi|_{\partial D} = 0$. Given the velocity u , we recover the stream function by means of the vorticity $\omega = \text{curl } u = -\Delta \psi$, so $\psi(x) = \int_D g(x, y) \omega(y) dy$ (here g is the Green's function of the Laplacian $-\Delta$ and x, y are points in D). The Euler equations can be written as

$$\begin{aligned} \frac{\partial \omega}{\partial t} + u \cdot \nabla \omega &= 0 \\ \omega &= \text{curl } u \end{aligned} \quad [31]$$

Consider now a solution given by vorticity concentrated in a finite number N of points:

$$\omega = \sum_{i=1}^N \lambda_i \delta_{x_i(t)} \quad [32]$$

Here the vortex intensities λ_i are real values and $x_i(t)$ are distinct points in D for $i=1, \dots, N$.

According to the Euler equations, these points evolve as follows (see also Marchioro and Pulvirenti (1994)):

$$\begin{aligned} \frac{d}{dt} x_j(t) &= \nabla_{x_j}^\perp \sum_{l=1, l \neq j}^N \lambda_l g(x_j(t), x_l(t)) \\ &+ \lambda_j \nabla_{x_j}^\perp \tilde{g}(x_j), \quad j = 1, \dots, N \end{aligned} \quad [33]$$

where \tilde{g} is related to the Green's function g . This is a Hamiltonian system in D^N . Hereafter, we shall suppose that the vortex intensities are the same ($\lambda_i = \lambda \forall i$), so that the Hamiltonian is

$$\mathcal{H}(x_1, \dots, x_N) = \frac{1}{2} \sum_{l,j=1, l \neq j}^N g(x_j, x_l) + \sum_{j=1}^N \tilde{g}(x_j) \quad [34]$$

By means of \mathcal{H} , we define the canonical Gibbs measure

$$\begin{aligned} \mu^N(dx_1 dx_2 \cdots dx_N) \\ = \frac{1}{Z(N)} e^{-\tilde{\beta} \lambda \mathcal{H}(x_1, \dots, x_N)} dx_1 dx_2 \cdots dx_N \end{aligned} \quad [35]$$

where $Z(N)$ is the partition function. If $Z(N) < \infty$, then μ^N is a well-defined probability measure on D^N and, by construction, it is an invariant measure for

system [33]. We can prove that $Z(N)$ is finite for $\tilde{\beta} \lambda \in (-8\pi/N, 4\pi)$, so that it is natural to choose as a scaling $\tilde{\beta} \lambda N = \beta$. Hence,

$$\begin{aligned} \mu^N(dx_1 dx_2 \cdots dx_N) \\ = \frac{1}{Z(N)} e^{-(\beta/N) \mathcal{H}} dx_1 dx_2 \cdots dx_N \end{aligned} \quad [36]$$

is considered for $-8\pi < \beta \leq 0$, or $\beta > 0$ with $N > \beta/4\pi$.

Bearing in mind the Onsager approach to approximate the turbulent Euler motion by means of point vortices, we are interested in the limit as N goes to $+\infty$, for β fixed in $(-8\pi, +\infty)$. It turns out that, when the number of point vortices becomes very large, their statistical behavior corresponds to a very large number of independent particles moving in a mean force field that they create.

More precisely, consider $\lambda = 1/N, \tilde{\beta} = \beta$. The empirical measure

$$\frac{1}{N} \sum_{i=1}^N \delta_{x_i(t)}$$

describing the vorticity, weakly converges to a probability density ρ and each correlation function

$$\begin{aligned} \rho_j^N(x_1, \dots, x_j) &= \int_D dx_{j+1} \cdots \int_D dx_N \frac{1}{Z(N)} e^{-(\beta/N) \mathcal{H}} \\ &\text{for } j = 1, \dots, N-1 \end{aligned} \quad [37]$$

weakly converges to $\otimes_{i=1}^j \rho = \prod_{i=1}^j \rho(x_i)$.

The equation satisfied by ρ , also called the mean-field equation, is

$$\begin{aligned} \rho(x) &= \frac{e^{-\beta U(x)}}{\int_D e^{-\beta U(y)} dy}, \\ &\text{with } U(x) = \int_D g(x, y) \rho(y) dy \end{aligned} \quad [38]$$

The relation between U and ρ can also be written as $-\Delta U = \rho$ in D , $U = 0$ on ∂D . We point out that $u = \nabla^\perp U$ is a stationary solution of the Euler equations. Indeed, $\omega = -\Delta U = \rho$ and ρ is a function of U , let us say $\rho = F(U)$. This gives that $\nabla \omega = \nabla U F'(U)$ and thus the term $u \cdot \nabla \omega$ in the Euler equation [31] vanishes.

It can be proved that there exists a solution of the mean-field equation when $\beta \geq 0$ or when $\beta < 0$ and D is simply connected. Uniqueness is known in some cases, for instance, when D is a bounded open smooth simply connected domain and the velocity is assumed tangent to the boundary.

There are numerical evidences of this approximation approach (see references in Lions (1997) referring to the periodic case). They show that for

large time and large Reynolds number (viscosity ν close to 0), the vorticity of the solution of the Navier–Stokes equations appears in a simple and organized structure. This stays intact until the viscous dissipation damps it. The important observation is that the organized structure is described quite precisely by the solution of the mean-field equation for some specific β .

Actually, to say that a fluid is inviscid is an approximation (which may be justified in many contexts), since every fluid displays some kind of viscosity. But turbulence is a phenomenon occurring at very small viscosity. In this sense, the above result provides a description of stationary regime in an ideal fluid, which is a good approximation of some numerical simulations of real fluids. Besides this good agreement with numerical simulations, there is no proof on how to deduce the mean-field equation from the Euler equations (e.g., which parameter β has to be chosen in eqn [38]?).

Remark The extension of this analysis to three-dimensional flows involves vortex filaments, instead of point vortices. There are attempts to describe interacting vortex filaments as proposed by Chorin. Idealizations of behavior of vortices are introduced to have a tractable mathematical model. The reader is referred to Lions (1997) for a description of nearly parallel vortex filaments and to Flandoli and Bessaih (2003) for more realistic filaments which fold.

See also: Cauchy Problem for Burgers-Type Equations; Hamiltonian Fluid Dynamics; Incompressible Euler Equations: Mathematical Theory; Malliavin Calculus; Non-Newtonian Fluids; Partial Differential Equations: Some Examples; Stochastic Differential Equations; Turbulence Theories; Viscous Incompressible Fluids: Mathematical Theory; Vortex Dynamics.

Further Reading

Albeverio S and Cruzeiro AB (1990) Global flows with invariant (Gibbs) measures for Euler and Navier–Stokes two dimensional fluids. *Communications in Mathematical Physics* 129: 431–444.
 Bensoussan A and Temam R (1973) Équations stochastiques du type Navier–Stokes. *Journal of Functional Analysis* 13: 195–222.
 Bricmont J, Kupiainen A, and Lefevere R (2001) Ergodicity of the 2D Navier–Stokes equations with random forcing. *Communications in Mathematical Physics* 224(1): 65–81.

Da Prato G and Debussche A (2003) Ergodicity for the 3D stochastic Navier–Stokes equations. *Journal de Mathématiques Pures et Appliquées* 82(8): 877–947.
 E Weinan, Mattingly JC, and Sinai Y (2001) Gibbsian dynamics and ergodicity for the stochastically forced Navier–Stokes equation. *Communications in Mathematical Physics* 224(1): 83–106.
 Flandoli F and Bessaih H (2003) A mean field result for 3D vortex filaments. In: Davies IM, Jacob N, Truman A, Hassan O, Morgan K, and Weatherill NP (eds.) *Probabilistic Methods in Fluids*, 22–34. River Edge, NJ: World Scientific.
 Flandoli F and Maslowski B (1995) Ergodicity of the 2-D Navier–Stokes equation under random perturbations. *Communications in Mathematical Physics* 172(1): 119–141.
 Frisch U (1995) *Turbulence. The legacy of A. N. Kolmogorov*. Cambridge: Cambridge University Press.
 Gallavotti G (2002) *Foundations of Fluid Dynamics*. Berlin: Springer.
 Hairer M and Mattingly JC (2004) Ergodic properties of highly degenerate 2D stochastic Navier–Stokes equations (English. English, French summary). *Comptes Rendus Mathématique. Académie des Sciences. Paris* 339(12): 879–888 (see also the paper “Ergodicity of the 2D Navier–Stokes equations with degenerate stochastic forcing” to appear on *Annals of Mathematics*).
 Hopf E (1952) Statistical hydromechanics and functional calculus. *Journal of Rational Mechanics and Analysis* 1: 87–123.
 Kraichnan RH and Montgomery D (1980) Two-dimensional turbulence. *Reports on Progress in Physics* 43(5): 547–619.
 Kuksin SB (2004) The Eulerian limit for 2D statistical hydrodynamics. *Journal of Statistical Physics* 115(1/2): 469–492.
 Kuksin S and Shirikyan A (2000) Stochastic dissipative PDEs and Gibbs measures. *Communications in Mathematical Physics* 213(2): 291–330.
 Lions P-L (1997) *On Euler Equations and Statistical Physics*, Pubblicazione della Scuola Normale Superiore. Pisa: Cattedra Galileiana.
 Marchioro C and Pulvirenti M (1994) Mathematical Theory of Incompressible Nonviscous Fluids. *Applied Mathematical Science*, vol. 96. New York: Springer.
 Mikulevicius R and Rozovskii BL (2004) Stochastic Navier–Stokes equations for turbulent flows. *SIAM Journal on Mathematical Analysis* 35(5): 1250–1310.
 Monin AS and Yaglom AM (1987) *Statistical Fluid Mechanics: Mechanics of Turbulence*. Cambridge, MA: MIT Press.
 Onsager L (1949) Statistical hydrodynamics. *Nuovo Cimento* 6(suppl. 2): 279–287.
 Robert R (2003) Statistical hydrodynamics (Onsager revisited). In: Friedlander S and Serre D (eds.) *Handbook of Mathematical Fluid Dynamics*, vol. II, pp. 1–54. Amsterdam: North-Holland.
 Vishik MJ and Fursikov AV (1988) *Mathematical Problems of Statistical Hydromechanics*. Dordrecht: Kluwer Academic Publishers.

Stochastic Loewner Evolutions

G F Lawler, Cornell University, Ithaca, NY, USA

© 2006 Gregory F Lawler. Published by Elsevier Ltd.

All rights reserved.

Introduction

The stochastic Loewner evolution or Schramm–Loewner evolution (SLE) is a family of random curves that appear as scaling limits of curves or cluster boundaries of discrete statistical mechanical models in two dimensions at criticality. The stochastic Loewner evolution was introduced by Oded Schramm as a candidate for the limit of loop-erased random walk and the boundary of percolation clusters, and it is now believed that SLE curves appear in most planar critical systems whose scaling limit satisfies conformal invariance. The curves are defined by solving a Loewner differential equation with a random input.

Definition

There are three major one-parameter families of SLE curves – chordal, radial, and whole-plane – which correspond to curves connecting two boundary points in a domain, a boundary point and an interior point in a domain, and two points in \mathbb{C} , respectively. The parameter is usually denoted $\kappa > 0$. The starting point for defining SLE is to write down the assumptions that one expects from a scaling limit, assuming that the limit is conformally invariant.

In the chordal case, we assume that there is a family of probability measures $\{\mu_D(z, w)\}$, indexed by simply connected proper domains $D \subset \mathbb{C}$ and distinct boundary points $z, w \in \partial D$, supported on continuous curves $\gamma: [0, t_\gamma] \rightarrow \bar{D}$ with $\gamma(0) = z$, $\gamma(t_\gamma) = w$, which satisfies the following:

- *Conformal invariance.* If $f: D \rightarrow D'$ is a conformal transformation, then the image of $\mu_D(z, w)$ under f is the same as $\mu_{D'}(f(z), f(w))$, up to a time change.
- *Conformal Markov property for $\mu_D(z, w)$.* Suppose $\gamma[0, t]$ is known, and let g_t be a conformal transformation of the slit domain $D \setminus \gamma[0, t]$ onto D with $g_t(\gamma(t)) = z$, $g_t(w) = w$ (see Figure 1). Then the conditional distribution on $g_t \circ \gamma[t, t_\gamma]$, given $\gamma[0, t]$, is the same, up to a change of parametrization, as the original distribution. (Implicit in this is the assumption that $\gamma(t)$ is on the boundary of $D \setminus \gamma(0, t]$, which will be true, e.g., if γ is non-self-intersecting and $\gamma(0, t_\gamma) \subset D$.)

Using the Riemann mapping theorem, one can see that such a family $\{\mu_D(z, w)\}$ is determined (up to reparametrization) by $\mu_{\mathbb{H}}(0, \infty)$, where $\mathbb{H} = \{x + iy: y > 0\}$ denotes the upper half-plane. Suppose $\gamma: [0, \infty) \rightarrow \mathbb{C}$ is a simple (i.e., no self-intersections) curve with $\gamma(0) = 0$, $\gamma(0, \infty) \subset \mathbb{H}$, and $\sup_t \text{Im}[\gamma(t)] = \infty$. Let $H_t = \mathbb{H} \setminus \gamma[0, t]$. There is a unique conformal transformation $g_t: H_t \rightarrow \mathbb{H}$ whose expansion at infinity is

$$g_t(z) = z + \frac{b(t)}{z} + O(|z|^{-2}), \quad z \rightarrow \infty$$

(see Figure 2). The coefficient $b(t)$, which is sometimes called the half-plane capacity of $\gamma[0, t]$ and denoted $\text{hcap}[\gamma[0, t]]$, is continuous, strictly increasing, and tending to ∞ . In fact,

$$b(t) = \lim_{y \rightarrow \infty} y E[\text{Im}[X_\tau] \mid X_0 = iy]$$

where X_s denotes a complex Brownian motion and $\tau = \tau_{\gamma[0, t]}$ is the first time s such that $X_s \in \mathbb{R} \cup \gamma[0, t]$. By reparametrizing γ , $b(t) = 2t$. With this parametrization, the maps g_t satisfy the Loewner differential equation

$$\dot{g}_t(z) = \frac{2}{g_t(z) - U_t}, \quad g_0(z) = z$$

where $U: [0, \infty) \rightarrow \mathbb{R}$ is a continuous function with $U_0 = 0$. In fact, $U_t = g_t(\gamma(t))$. Schramm observed that the measure $\mu_{\mathbb{H}}(0, \infty)$, at least if it were supported on simple curves and the curves were parametrized using half-plane capacity, would produce a random U_t . If the assumptions above on $\{\mu_D(z, w)\}$ are translated into assumptions on the “driving function” U_t , one shows readily that U_t must be a driftless Brownian motion, that is, $U_t = \sqrt{\kappa} B_t$, for a standard one-dimensional Brownian motion B_t .

Chordal SLE $_{\kappa}$ (in \mathbb{H} connecting 0 and ∞) is defined to be the random collection of conformal maps g_t obtained by solving the initial-value problem

$$\dot{g}_t(z) = \frac{2}{g_t(z) - \sqrt{\kappa} B_t}, \quad g_0(z) = z \tag{1}$$

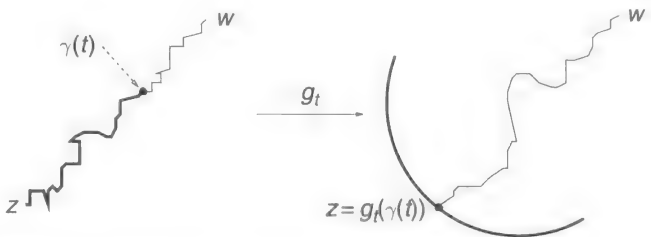


Figure 1 The map g_t from $D \setminus \gamma[0, t]$ onto D .

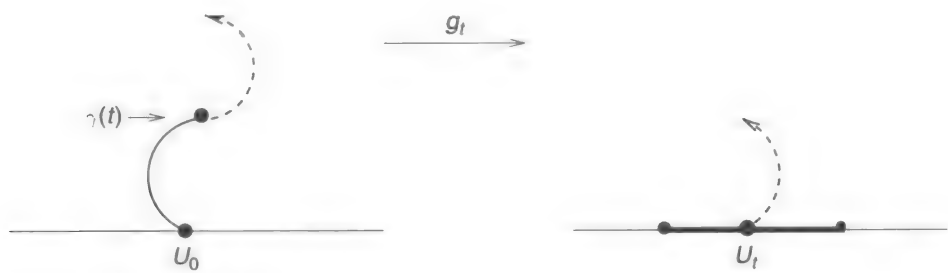


Figure 2 The map g_t from $\mathbb{H} \setminus \gamma[0, t]$ onto \mathbb{H} .

where B_t is a standard one-dimensional Brownian motion. Equation [1] is often given in terms of the inverse $f_t = g_t^{-1}$:

$$\dot{f}_t(z) = -f'_t(z) \frac{2}{z - \sqrt{\kappa} B_t}$$

This equation describes a random evolution of conformal maps f_t from \mathbb{H} into subdomains of \mathbb{H} . For each $z \in \mathbb{H}$, the solution of [1] is defined up to a time $T_z \in [0, \infty]$ with $T_z > 0$ for $z \neq 0$. For fixed t , g_t is the unique conformal transformation of $H_t := \{z \in \mathbb{H} : T_z > t\}$ onto \mathbb{H} with expansion

$$g_t(z) = z + \frac{2t}{z} + \dots, \quad z \rightarrow \infty$$

The chordal SLE_κ path is the random curve $\gamma : [0, \infty) \rightarrow \mathbb{H}$ such that for each t , H_t is the unbounded component of $\mathbb{H} \setminus \gamma[0, t]$. It is not immediate from the definition that such a curve γ exists, but its existence has been proved. If $G_t = g_{t/\kappa}$, then we can write eqn [1] as

$$\dot{G}_t(z) = \frac{a}{G_t(z) + W_t} \tag{2}$$

where $a = 2/\kappa$ and $W_t := -\sqrt{\kappa} B_{t/\kappa}$ is a standard Brownian motion. Then $Z_t^z := G_t(z) + W_t$ satisfies the Bessel stochastic differential equation

$$dZ_t^z = \frac{a}{Z_t^z} dt + dW_t, \quad Z_0^z = z \tag{3}$$

This equation is valid up to time κT_z , which is the first time that $Z_t^z = 0$.

Although chordal SLE_κ is defined with a particular parametrization, one generally thinks of it as a measure on curves modulo reparametrization. The scaling properties of Brownian motion imply that this measure is invariant under dilations of \mathbb{H} . If D is a simply connected domain and z, w are distinct boundary points of D , chordal SLE_κ in D connecting z and w is defined to be the conformal image of SLE_κ in \mathbb{H} from 0 to ∞ under a conformal transformation of \mathbb{H} onto D taking 0 to z and ∞ to w . There is a one-parameter family of such

transformations, but the scale invariance of SLE_κ in \mathbb{H} shows that the image measure is independent of the choice of transformation.

The geometric and fractal properties of the curve γ vary greatly as the parameter κ changes:

- if $\kappa \leq 4$, γ is a simple curve;
- if $4 < \kappa < 8$, γ has self-intersections, but is not space filling; and
- if $\kappa \geq 8$, γ is a space filling curve.

To see this, one notes that the conformal Markov property implies that there can be double points with positive probability if and only if $T_x < \infty$ occurs with positive probability for $x > 0$. In addition, the curve is space filling if and only if $T_z < \infty$ for all z and $T_w \neq T_z$ for $w \neq z$. The problem is then reduced to a problem about the Bessel equation [3] for which the following holds:

- if $a \geq 1/2$ and $z \neq 0$, the probability that $T_z < \infty$ is zero. If $a < 1/2$, this probability equals 1.
- if $1/4 < a < 1/2$, and w, z are distinct points in \mathbb{H} , then there is a positive probability that $T_w = T_z$.
- if $0 < a \leq 1/4$, then with probability 1, $T_w \neq T_z$ for all $w \neq z$.

This kind of argument is typical when studying SLE – geometric properties of the curve are established by analyzing a stochastic differential equation. The Hausdorff dimension of the path γ is given by

$$\dim[\gamma[0, \infty)] = \min\left\{1 + \frac{\kappa}{8}, 2\right\}$$

The radial Loewner equation describes the evolution of a curve from the boundary of the unit disk $\mathbb{D} = \{z : |z| < 1\}$ to the origin. Suppose $\gamma : [0, \infty) \rightarrow \bar{\mathbb{D}}$ is a simple curve with $\gamma(0) = 1$, $\gamma(0, \infty) \subset \mathbb{D} \setminus \{0\}$, and $\gamma(t) \rightarrow 0$ as $t \rightarrow \infty$. Let g_t be the unique conformal transformation of $\mathbb{D} \setminus \gamma[0, t]$ onto \mathbb{D} such that $g_t(0) = 0$, $g'_t(0) > 0$. One can check that $g'_t(0)$ is continuous and strictly increasing in t , and hence we can parametrize γ in such a way that $g'_t(0) = e^t$. Using this reparametrization, there is a continuous

$U_t: [0, \infty) \rightarrow \mathbb{R}$ with $U_0 = 0$ such that g_t satisfies the radial Loewner equation

$$\dot{g}_t(z) = g_t(z) \frac{e^{iU_t} + g_t(z)}{e^{iU_t} - g_t(z)}, \quad g_0(z) = z$$

If $z \neq 0$, then we can define $h_t(z) = -i \log g_t(z)$ locally near z , and this equation becomes

$$\dot{h}_t(z) = \cot\left(\frac{h_t(z) - U_t}{2}\right)$$

Radial SLE_κ (connecting 1 and 0 in \mathbb{D}) is obtained by setting $U_t = \sqrt{\kappa} B_t$. If D is a simply connected domain, $z \in D, w \in \partial D$, then radial SLE_κ in D connecting w and z is obtained by conformal transformation using the unique transformation f of \mathbb{D} onto D with $f(0) = z, f(1) = w$. Again, we think of this as being defined modulo time change. If $a = 2/\kappa$ and $v_t = h_{at/2}$, then

$$\dot{v}_t(z) = \frac{a}{2} \cot\left(\frac{v_t(z) + W_t}{2}\right) \quad [4]$$

where $W_t := -\sqrt{\kappa} B_{t/\kappa}$ is a standard Brownian motion. If $L_t^z = v_t(z) + W_t$, then we get

$$dL_t^z = \frac{a}{2} \cot\left(\frac{L_t^z}{2}\right) dt + dW_t$$

Radial and chordal SLE are closely related. In fact, if γ is a chordal SLE path in \mathbb{H} from 0 to ∞ , $\tilde{\gamma}$ is a radial SLE path in \mathbb{D} from 1 to 0, and $\eta = -i \log \tilde{\gamma}$, then for small t the distribution of η is absolutely continuous to the distribution of a (random time change of) γ . Showing this involves understanding the behavior of the Loewner equation under conformal transformations. Suppose $\gamma, \tilde{\gamma}$ have been parametrized as in [2] and [4] with $a = 2/\kappa$. Let g_t^* be the conformal transformation of $\mathbb{H} \setminus \eta[0, t]$ onto \mathbb{H} such that

$$g_t^*(z) = z + \frac{a^*(t)}{z} + \dots, \quad z \rightarrow \infty$$

and let U_t^* be the Loewner driving function such that

$$\dot{g}_t^*(z) = \frac{\dot{a}^*(t)}{g_t^*(z) - U_t^*}$$

Here $a^*(t) = \text{hcap}[\eta[0, t]]$. If we consider a time change σ such that $a^*(\sigma(t)) = at$ and let $U_t = \tilde{U}_{\sigma(t)}$ be the time-changed driving function, Itô's formula can be used to show that

$$dU_t = \frac{1}{2}(1 - 3a) F_t dt + d\tilde{W}_t \quad [5]$$

where the F_t in the drift term depends on $\gamma[0, t]$ and is independent of a , and \tilde{W} is a standard Brownian

motion. Girsanov's theorem implies that Brownian motions with the same variance but different drifts have absolutely continuous distributions. In particular, qualitative properties such as existence of double points or Hausdorff dimension of paths are the same for radial and chordal SLE. U_t is a driftless Brownian motion if $a = 1/3, \kappa = 6$.

Whole-plane SLE_κ from 0 to ∞ is a path $\gamma: (-\infty, \infty) \rightarrow \mathbb{C}$ with $\gamma(-\infty) = 0, \gamma(\infty) = \infty$, such that given $\gamma(-\infty, t]$, the distribution of $\gamma(t, \infty)$ is that of radial SLE_κ from boundary point $\gamma(t)$ to interior point ∞ in the domain $\mathbb{C} \setminus \gamma[-\infty, t]$. One can define whole-plane SLE_κ connecting two distinct points in \mathbb{C} by conformal transformation.

Locality and Restriction

There are two special values of κ : $\kappa = 6, a = 1/3$ that satisfies the "locality" property and $\kappa = 8/3, a = 3/4$ that satisfies the "restriction" property. Suppose γ is a chordal SLE_κ curve from 0 to ∞ in \mathbb{H} parametrized as in [2]. Suppose $\Phi: \mathcal{N} \rightarrow H$ is a conformal map taking a neighborhood \mathcal{N} of 0 in \mathbb{H} to $\Phi(\mathcal{N})$ and that locally maps \mathbb{R} into \mathbb{R} . Let $\tilde{\gamma}(t) = \Phi \circ \gamma(t)$, which is defined for sufficiently small t . Let g_t^* be the conformal transformation of $\mathbb{H} \setminus \tilde{\gamma}[0, t]$ onto \mathbb{H} with

$$g_t^*(z) = z + \frac{a^*(t)}{z} + \dots$$

and let \tilde{U}_t be the driving function such that

$$\dot{g}_t^*(z) = \frac{\dot{a}^*(t)}{g_t^*(z) - \tilde{U}_t}$$

Here $a^*(t) = \text{hcap}[\tilde{\gamma}[0, t]]$. If we change time, $\gamma_t^* = \tilde{\gamma}_{\sigma(t)}$, so that $a^*(\sigma(t)) = at$, then an application of Itô's formula shows that $U_t^* := \tilde{U}_{\sigma(t)}$ satisfies

$$dU_t^* = \frac{1}{2}(3a - 1) \frac{\Phi''_{\sigma(t)}(W_{\sigma(t)})}{\Phi'_{\sigma(t)}(W_{\sigma(t)})^2} dt + d\tilde{W}_t$$

Here \tilde{W}_t is a standard Brownian motion, $\Phi_t = g_t^* \circ \Phi \circ g_t^{-1}$, and g_t is the conformal map associated to γ . In particular, if $a = 1/3, \kappa = 6, U_t^*$ is a standard Brownian motion; hence, $\tilde{\gamma}^*$ has the distribution of SLE_6 . The locality property for SLE_6 can be stated as "the conformal image of SLE_6 is (a time change of) SLE_6 ." Intuitively, the SLE_6 path in a restricted domain does not feel the boundary of the domain until it reaches it. Radial SLE_6 satisfies a similar locality property. Moreover, [5] can be used to show that the image of chordal SLE_6 under the exponential map is the same (for small time t) as radial SLE_6 . The locality property explains why

SLE_6 is a natural candidate for the boundary of percolation clusters.

If $\kappa \leq 4$, SLE_κ paths are simple, that is, with no self-intersections. Suppose $A \subset \bar{\mathbb{H}} \setminus \{0\}$ is a compact set such that $\mathbb{H} \setminus A$ is simply connected. Let γ denote a chordal SLE_κ in \mathbb{H} connecting 0 and ∞ and let E_A be the event $E_A = \{\gamma(0, \infty) \cap A = \emptyset\}$. Let $\Phi_A: \mathbb{H} \setminus A \rightarrow \mathbb{H}$ be the unique conformal transformation with $\Phi_A(0) = 0$, $\Phi_A(\infty) = \infty$, $\Phi'_A(\infty) = 1$. On the event E_A , we can define $\tilde{\gamma}(t) = \Phi_A \circ \gamma(t)$. Chordal SLE_κ is said to satisfy the restriction property if the conditional distribution of $\tilde{\gamma}$ given E_A is the same as (a time change of) γ . The only $\kappa \leq 4$ that satisfies this property is $\kappa = 8/3$. The proof of this fact also establishes the formula: if γ is a chordal $SLE_{8/3}$ curve in \mathbb{H} from 0 to ∞ , then

$$P\{\gamma(0, \infty) \cap A = \emptyset\} = \Phi'_A(0)^{5/8} \quad [6]$$

There is a similar formula for radial $SLE_{8/3}$, which establishes a radial restriction property. Suppose $A \subset \bar{\mathbb{D}} \setminus \{0, 1\}$ is a compact set such that $\mathbb{D} \setminus A$ is simply connected. Let Ψ_A be the unique conformal transformation of $\mathbb{D} \setminus A$ onto \mathbb{D} with $\Psi_A(0), \Psi'_A(0) > 0$. Then, if γ is a radial $SLE_{8/3}$ curve from 1 to 0 in \mathbb{D} , then

$$P\{\gamma(0, \infty) \cap A = \emptyset\} = \Psi'_A(0)^{5/48} |\Psi'_A(1)|^{5/8}$$

The restriction property makes $SLE_{8/3}$ the candidate for the scaling limit of self-avoiding walks.

Relation to Conformal Field Theory

The Schramm–Loewner evolution is one of the tools used to rigorously prove predictions made using powerful, yet nonrigorous, arguments of conformal field theory. In conformal field theory, there is a parameter c , called the central charge, which classifies theories. To each $c \leq 1$, there corresponds a $\kappa \leq 4$ and a “dual” $\kappa' = 16/\kappa \geq 4$:

$$c = c_\kappa = \frac{(8 - 3\kappa)(\kappa - 6)}{2\kappa}$$

In particular, $\kappa = 8/3$, $\kappa' = 6$ corresponds to central charge zero. It is expected, and has been proved in a number of cases, that SLE_κ or $SLE_{\kappa'}$ curves will appear in scaling limits of systems with central charge c_κ . These systems can also be parametrized by the boundary scaling exponent or conformal weight

$$\alpha = \alpha_\kappa = \frac{6 - \kappa}{2\kappa}$$

For $\kappa = 8/3$, $\alpha = 5/8$ which is the exponent in [6].

In studying the relationship between SLE and conformal field theories, two other probabilistic objects, restriction measures and the (Brownian) loop soup, arise. An \mathbb{H} -hull (connecting 0 and ∞) is an unbounded, connected, closed set $K \subset \bar{\mathbb{H}}$ with $K \cap \mathbb{R} = \{0\}$ and such that $\mathbb{H} \setminus K$ consists of two connected components, one whose boundary includes the positive reals and the other whose boundary includes the negative reals. A (chordal) restriction measure on hulls K is a probability measure with the property that for any A as in [6], the distribution of $\Phi_A \circ K$ given $\{K \cap A = \emptyset\}$ is the same as the original measure. The (Brownian) loop measure is a measure on unrooted loops derived from Brownian bridges. It is the scaling limit of the measure on random walk loops that gives each unrooted simple random walk loop of length $2n$ measure 4^{-2n} . The loop measure in a bounded domain is obtained by restricting to loops that stay in that domain. We can consider this as a measure on “hulls” by filling in the bounded holes (so that the complement of the hull is connected). By doing this we get a family of infinite measures on hulls, indexed by domains D , and this family satisfies conformal invariance and the restriction property. The loop soup with parameter λ is a Poissonian realization from this measure with parameter λ .

The set of all restriction measures is parametrized by $\alpha \geq 5/8$; the α -restriction measure has the property that

$$P\{K \cap A \neq \emptyset\} = \Phi'_A(0)^\alpha$$

For $\alpha = 5/8$, K is given by the path of $SLE_{8/3}$. For integer α , the hull K can be constructed by taking α -independent Brownian excursions in \mathbb{H} (Brownian motions starting at 0 conditioned to stay in \mathbb{H} for all times), and letting K be the hull obtained by taking the union of the paths and filling in the bounded holes. If $\kappa \leq 8/3$, $c_\kappa \leq 0$, then the restriction measure with exponent $\alpha_\kappa \geq 5/8$ can also be constructed as follows: take a chordal SLE_κ path and an independent realization of the loop soup with intensity $\lambda_\kappa = -c_\kappa$; add to the SLE path all the loops in the soup that intersect the SLE_κ curve; and then fill in all the bounded hulls. The limiting case $\alpha = 5/8$, $\lambda = 0$ gives the only measure supported on simple curves that is also a restriction measure, $SLE_{8/3}$.

For $8/3 < \kappa \leq 4$, $0 < c_\kappa \leq 1$, it is conjectured, and proved for small c_κ , that SLE_κ curves can be found by taking a loop soup with parameter $\lambda = c_\kappa$ and looking at connected curves in the fractal set given by the complement of the union of all the hulls generated by the loops.

Examples

The scaling limit of simple random walk, Brownian motion, is known to be conformally invariant. A two-dimensional Brownian bridge or loop is a Brownian motion, $B_t, 0 \leq t \leq 1$, conditioned so that $B_0 = B_1$. The frontier or outer boundary of the Brownian motion is the boundary of the unbounded component of the complement. Benoit Mandelbrot first observed numerically that the outer boundary of Brownian motion had fractal dimension $\sim 4/3$. Gregory Lawler, Oded Schramm, and Wendelin Werner used SLE to prove that the boundary has Hausdorff dimension $4/3$. In fact, the outer boundary can be considered as an $\text{SLE}_{8/3}$ loop.

SLE_6 and $\text{SLE}_{8/3}$ arise in the scaling limit of critical percolation on the triangular lattice. Suppose that each vertex in the upper half-plane triangular lattice is colored black or white each with a probability $1/2$. Suppose the real line gives a boundary condition of black on the negative real line and white on the positive real line. Then if we represent the vertices in the lattice as hexagons as in the figure, a curve is formed which is the boundary between the black and white clusters. This curve is called the “percolation exploration process.” Stanislav Smirnov proved that the scaling limit of this curve is conformally invariant, and from this it can be concluded that the curve is chordal SLE_6 . In particular, the Hausdorff dimension is $7/4$ and the scaling limit has double points. In the scaling limit, the “outer boundary” of this curve has Hausdorff dimension $4/3$ and its dimension is absolutely continuous with respect to that of $\text{SLE}_{8/3}$. While this result is expected for other critical percolation model, such as bond percolation in \mathbb{Z}^2 with critical probability $1/2$, it has only been proved for the triangular lattice. Percolation has central charge 0 and the “locality” property can be seen in the lattice model. The outer boundary of the curve has the same distribution as the outer boundary of a Brownian motion that is reflected at angle $\pi/3$ off the real line. Locally, the outer boundary of percolation, the outer boundary of complex Brownian motion, and $\text{SLE}_{8/3}$ all look the same, and it is expected that this will also be true for the scaling limit of self-avoiding walks.

There are three models derived in some way from simple random walk that have been proved to have scaling limits of SLE_κ . The loop-erased random walk (LERW) in a finite subset V of \mathbb{Z}^2 connecting two distinct points is obtained by taking a simple random walk from one point to the other and erasing loops chronologically. The LERW is closely related to uniform spanning trees; in fact, if one

chooses a spanning tree of V from the uniform distribution on all spanning trees, then the distribution of the unique path connecting the two points is exactly that of the LERW (see Figure 3). Another description of the LERW is as the Laplacian random walk: the LERW from z to w in V chooses a new step weighted by the value of the function that is harmonic on the complement of w and the path up to that point with boundary values 0 on the path and 1 on w . The LERW in the discrete upper half-plane can be obtained by erasing loops from a simple random walk excursion. The LERW and the uniform spanning tree are systems with central charge $c = -2$. It has been proved that the scaling limit of the LERW is SLE_2 ; hence, the paths have Hausdorff dimension $5/8$.

There is another path associated to spanning trees given by the one-to-one correspondence between spanning trees and Hamiltonian walks on a corresponding directed (Manhattan) lattice on the dual graph (see Figure 4). If the spanning trees, or equivalently the Hamiltonian walks, are chosen using the uniform distribution, then the scaling limit of this walk is the space-filling curve SLE_8 . Note that 2 and 8 are the dual values of κ associated to $c = -2$.

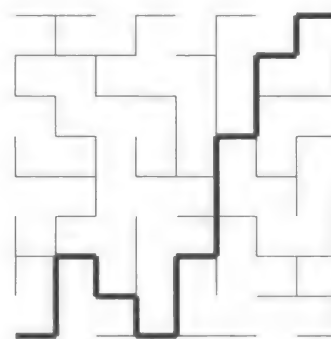


Figure 3 A spanning tree and the path between two vertices. If the tree has the uniform distribution, the path has the distribution of the LERW.

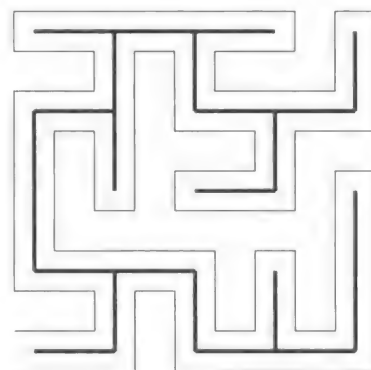


Figure 4 A spanning tree and the corresponding Hamiltonian walk.

Another discrete process derived from simple random walk, the harmonic explorer, has a scaling limit of SLE_4 . There is a particular property of SLE_κ that leads to the definition of this discrete process. Consider a chordal SLE_κ curve, let $z \in \mathbb{H}$, and let Z_t^z be as in [3] with $a = 2/\kappa$. Itô's formula shows that $\Theta_t := \arg(Z_t^z)$ satisfies

$$d\Theta_t = \left(\frac{1}{2} - a\right) \frac{\sin(2\Theta_t)}{|Z_t^z|^2} dt - \frac{\sin \Theta_t}{|Z_t^z|} dW_t$$

In particular, Θ_t is a martingale if and only if $a = 1/2$, $\kappa = 4$. The probability that a complex Brownian motion starting at $z \in \mathbb{H}$ first hits \mathbb{R} on the negative half-line can be shown to be $\arg(z)$. If $\kappa \leq 4$, then we can see that Θ_∞ equals 0 or π , depending on whether z is on the right or left side of the path $\gamma(0, \infty)$. For the martingale case $\kappa = 4$, Θ_t represents the probability that z is on the left side of $\gamma(0, \infty)$, given $\gamma(0, t]$. The harmonic explorer is a process on the hexagonal lattice defined to have this property. In a way similar to the percolation process, the walk is defined as the boundary between black and white hexagons on the triangular lattice. However, when an unexplored hexagon is reached in the harmonic explorer, it is colored black with probability q , where q is the probability that a simple random walk on the triangular lattice starting at that hexagon (considered as a vertex in the triangular lattice) hits a black hexagon before hitting a white hexagon. It is not difficult to show that this process has the property that for z away from the curve, the “probability of z ending on the left given the curve of n steps” is a martingale.

There are many other models for which SLE_κ curves are expected in the limit, but it has not been established. The most difficult part is to show the existence of a limit that is conformally invariant. One example is the self-avoiding walk (SAW). It is an open problem to establish that there exists a scaling limit of the uniform measure on SAWs and to establish conformal invariance of the limit. However, the nature of the discrete model is such that if the limit exists, it must satisfy the restriction property. Hence, under the assumption of conformal invariance, the only possible limit is $\text{SLE}_{8/3}$. Numerical simulations strongly support the conjecture that $\text{SLE}_{8/3}$ is the limit of SAWs, and this gives strong evidence for the conformal invariance conjecture for SAWs. Critical exponents for SAWs (as well as critical exponents for many other models) can be predicted nonrigorously from rigorous scaling exponents for the corresponding SLE paths.

Generalizations

One of the reasons that the theory of SLE is nice for simply connected domains is that a simply connected domain with an arc connected to the boundary of the domain removed is again simply connected. For nonsimply connected domains, it is more difficult to describe because the conformal type of the slit domain changes as time evolves. In the case of a curve crossing an annulus, this can be done with an added parameter referring to the conformal type of the annulus (two annuli of the form $\{z: r_1 < |z| < s_1\}$ are conformally equivalent if and only if $r_1/s_1 = r_2/s_2$). It is not immediately obvious what the correct definition of SLE should be in general domains and, more generally, on Riemann surfaces. One possibility for $\kappa \leq 4$ is to consider a configurational (equilibrium statistical mechanics) view of SLE. Consider a family of measures $\{\mu_D(z, w)\}$, where D ranges over domains and z, w are distinct boundary points at ∂D is locally analytic, supported on simple curves from z to w (modulo time change). Let $\mu_D^\#(z, w) = \mu_D(z, w)/|\mu_D(z, w)|$ be the corresponding probability measures, which may be defined even if ∂D is not smooth at z, w . Then the following axioms should hold:

- *Conformal invariance.* If $f: D \rightarrow D'$ is a conformal transformation, $f \circ \mu_D^\#(z, w) = \mu_{D'}^\#(f(z), f(w))$.
- *Conformal Markov property.*
- *Perturbation of domains.* Suppose $D_1 \subset D$ and $\partial D_1, \partial D$ agree near z, w . Then $\mu_{D_1}(z, w)$ should be absolutely continuous with respect to $\mu_D(z, w)$. Let Y denote the Radon–Nikodym derivative of $\mu_{D_1}(z, w)$ with respect to $\mu_D(z, w)$. Then

$$Y(\gamma) = 1\{\gamma(0, t_\gamma) \subset D_1\} F_c(D; \gamma, D \setminus D_1)$$

where F_c is to be determined. In the case where D, D_1 are simply connected, $F_c(D; \gamma, D \setminus D_1) = J(\gamma, D, D_1)^{-c}$, where $J(\gamma, D, D_1)$ denotes the probability that there is a loop in the Brownian loop soup in D that intersects both γ and $D \setminus D_1$. (There is no problem defining this quantity in nonsimply connected domains, but it is not clear that it is the right quantity.) Here $c = c_\kappa$. The restriction property tells us that $F_0 \equiv 1$.

- *Conformal covariance.* If f is as above, $\partial D, \partial D'$ are smooth near z, w and $f(z), f(w)$, respectively, then

$$f \circ \mu_D(z, w) = |f'(z)|^\alpha |f'(w)|^\alpha \mu_{D'}(f(z), f(w))$$

Here $\alpha = \alpha_\kappa$ is the boundary scaling exponent.

See also: Boundary Conformal Field Theory; Percolation Theory; Random Walks in Random Environments.

Further Reading

Kager W and Nienhuis B (2004) A guide to stochastic Loewner evolution and its applications. *Journal of Statistical Physics* 115: 1149–1229.

Lawler G (2004) Conformally invariant processes in the plane. In: *School and Conference on Probability Theory*, ICTP Lecture Notes, vol. 17, pp. 305–351. Trieste, Italy: ICTP.

Lawler G (2005) *Conformally Invariant Processes in the Plane*. Providence, RI: American Mathematical Society.

Lawler G and Werner W (2004) The Brownian loop soup. *Probability Theory Related Fields* 128: 565–588.

Lawler G, Schramm O, and Werner W (2004a) Conformal restriction, the chordal case. *The Journal of American Mathematical Society* 16: 917–955.

Lawler G, Schramm O, and Werner W (2004b) Conformal invariance of planar loop-erased random walks and uniform spanning trees. *Annals of Probability* 32: 939–995.

Schramm O (2000) Scaling limits of loop-erased random walks and uniform spanning trees. *Israel Journal of Mathematics* 118: 221–288.

Smirnov S (2001) Critical percolation in the plane: conformal invariance, Cardy’s formula, scaling limits. *Comptes rendus de l’Academie des sciences – series i – mathematics* 333: 239–244.

Werner W (2003) SLEs as boundaries of clusters of Brownian loops. *Comptes rendus de l’Academie des sciences – series i – mathematics* 337: 481–486.

Werner W (2004) Random planar curves and Schramm–Loewner evolutions. In: *Ecole d’Eté de Probabilités de Saint-Flour XXXII – 2002*, Lecture Notes in Mathematics, vol. 1840, pp. 113–195. Berlin: Springer.

Stochastic Resonance

S Herrmann, Université Henri Poincaré, Nancy 1
Vandœuvre-lès-Nancy, France
P Imkeller, Humboldt Universität zu Berlin,
Berlin, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The concept of stochastic resonance was introduced by physicists. It originated in a toy model designed for a qualitative description of periodic phenomena in the recurrences of glacial eras in Earth’s history. It spread its popularity over numerous areas of natural sciences: neuronal response to periodic stimuli, variations of magnetization in a ferromagnetic system, voltage variations in the simple Schmitt trigger electronic circuit or in more complicated devices, behavior of lasers in optical bi-stability, etc. The interest in this ubiquitous phenomenon is enhanced by signal analysis: an optimal dose of noise in some system can essentially boost signal transduction. Noise in this context does not enter the system as an impurity perturbing its performance, but on the contrary as a catalyst triggering amplified stochastic response to weak periodic signals.

The Climate Paradigm

The phenomenon of stochastic resonance was first discovered in an elementary climate model serving in an explanation of major transitions in paleoclimatic time series confining glacial cycles. Data collected for instance from ice or deep sea cores allow one to

deduce estimates of the average temperature on Earth over the last 700 000 years. They exhibit periodic switching between ice and warm ages with fast spontaneous transitions. The average periodicity of the glaciation time series obtained is $\sim 10^5$ years. In order to explain temperature variations, Benzi *et al.* (1981) introduced random perturbations into an energy balance model of the Budyko–Sellers type. This model describes the evolution of the seasonal and global average temperature X caused by defects in the balance between incoming and outgoing radiation

$$c \frac{dX(t)}{dt} = E_{in} - E_{out}$$

where c is the active thermal inertia of the system. The incoming energy is modeled as proportional to the “solar constant” Q :

$$E_{in} = Q \left(1 + A \cos \frac{2\pi t}{T} \right), \quad \text{with } T \approx 92\,000 \text{ years}$$

and $A \approx 0.1\%$ of Q . This exceedingly small variation of the solar constant is caused by a modulation of the orbital eccentricity of the Earth’s trajectory (Figure 1). The outgoing radiation E_{out} is composed

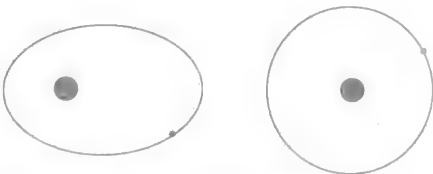


Figure 1 Modulation of the orbital eccentricity.

of two essential parts. The first part $a(X)E_{in}$ is dominated by the albedo $a(X)$ representing the proportion of energy reflected back to space. It is a decreasing function of temperature, due to the higher rate of reflection from a brighter Earth at low temperatures implying a bigger volume of ice. The second part of the outgoing radiation comes from the fact that the Earth radiates energy like a black body, and is given by the Boltzmann law γX^4 , where γ is the Stefan constant. Describing the balance of energy terms as a slowly and weakly time-varying gradient of a potential U , the balance model can be expressed by

$$\frac{dX(t)}{dt} = -\frac{\partial U}{\partial x}\left(\frac{t}{T}, X(t)\right)$$

where the time period 1 is blown up to (large) T by time scaling. The roles of deep and shallow wells switch periodically (Figure 2). Since the variation of the solar constant is extremely small, we can assume that the height of the barrier between the two wells is lower-bounded by a positive constant. The system then admits three steady states two of which are stable and separated by roughly 10 K. As the solar constant, they fluctuate slowly and very weakly. Therefore, this deterministic system cannot account for climate changes with temperature variations of ~ 10 K. They can only be explained by allowing transitions between the two steady states which become possible by adding noise to the system. In general, short timescale phenomena such as annual fluctuations in solar radiation are modeled by Gaussian white noise of intensity ε and lead to equations of the type

$$dX_t^\varepsilon = -\frac{\partial U}{\partial x}\left(\frac{t}{T}, X_t^\varepsilon\right) dt + \sqrt{\varepsilon} dW_t \quad [1]$$

which are generic for studying stochastic resonance in numerous physical and biological models. Generally, the input of noise amplifies a weak periodic signal by creating trajectories fluctuating randomly periodically between meta-stable states. An optimal tuning of noise intensity to period length (“stochastic resonance”) significantly enhances the response

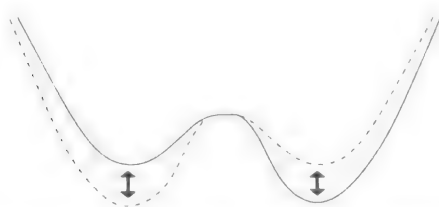


Figure 2 Deep and shallow wells switching periodically.

of the random system to weak perturbations with long periods.

Strongly Damped Brownian Particle

It is useful to roughly compare solutions of stochastic differential equations and motions of Brownian particles in double-well landscapes (Figure 3) in order to understand properties of their trajectories (see Schweitzer 2003, Mazo 2002). As in the previous section, let us concentrate on a one-dimensional setting, remarking that we shall give a treatment that easily generalizes to the finite-dimensional setting. Due to Newton’s law, the motion of a particle is governed by the impact of all forces acting on it. Let us denote F the sum of these forces, m the mass, x the space coordinate, and v the velocity of the particle. Then

$$m\dot{v} = F$$

Let us first assume the potential to be switched off. In their pioneering work at the turn of the twentieth century, Marian v. Smoluchowski and Paul Langevin introduced stochastic concepts to describe the Brownian particle motion by claiming that at time t

$$F(t) = -\gamma_0 v(t) + \sqrt{2k_B T \gamma_0} \dot{W}_t$$

The first term results from friction γ_0 and is velocity dependent. An additional stochastic force represents random interactions between Brownian particles and their simple molecular random environment. The white noise \dot{W} (formal derivative of the Wiener process) plays the crucial role. The diffusion coefficient (standard deviation of the random impact) is composed of Boltzmann’s constant k_B , friction, and environmental temperature T . It satisfies the condition of the fluctuation–dissipation theorem expressing the balance of energy loss due to friction and energy gain resulting from noise. The equation of motion becomes

$$\begin{aligned} \frac{dx(t)}{dt} &= v(t) \\ dv(t) &= -\frac{\gamma_0}{m} v(t) dt + \frac{\sqrt{2k_B T \gamma_0}}{m} dW_t \end{aligned}$$



Figure 3 Brownian particle in a double-well landscape.

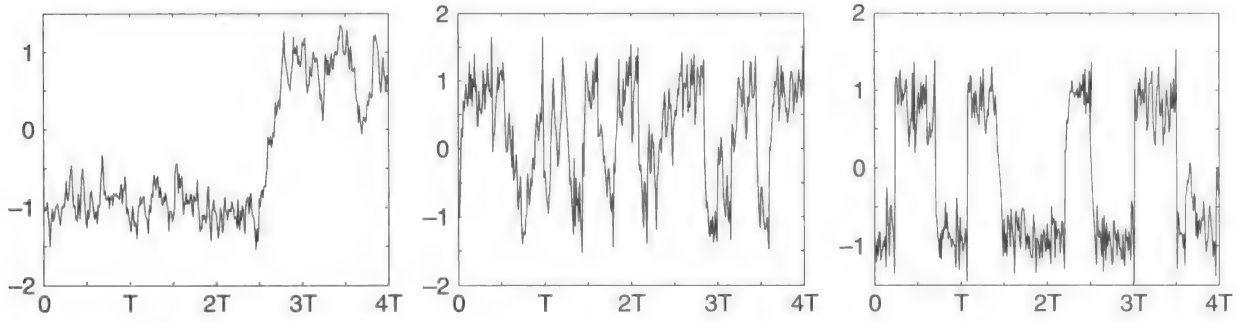


Figure 4 Resonance pictures for diffusions.

In the stationary regime, the stationary Ornstein–Uhlenbeck process provides its solution

$$v(t) = v(0) e^{-(\gamma_0/m)t} + \frac{\sqrt{2k_B T \gamma_0}}{m} \int_0^t e^{-(\gamma_0/m)(t-s)} dW_s$$

The ratio $\beta := \gamma_0/m$ determines the dynamic behavior. Let us focus on the over-damped situation with large friction and very small mass. Then for $t \gg 1/\beta = \tau$ (relaxation time), the first term in the expression for velocity can be neglected, while the stochastic integral represents a Gaussian process. By integrating, we obtain in the over-damped limit ($\beta \rightarrow \infty$) that v and thus x is Gaussian with almost constant mean

$$m(t) = x(0) + \frac{1 - e^{-\beta t}}{\beta} v(0) \approx x(0)$$

and covariance close to the covariance of white noise see Nelson (1967):

$$\begin{aligned} K(s, t) &= \frac{2k_B T}{\gamma_0} \min(s, t) + \frac{k_B T}{\gamma_0 \beta} (-2 + 2e^{-\beta t} + 2e^{-\beta s} \\ &\quad - e^{-\beta|t-s|} - e^{-\beta(t+s)}) \\ &\approx \frac{2k_B T}{\gamma_0} \min(s, t) \end{aligned}$$

Hence, the time-dependent change of the velocity of the Brownian particle can be neglected, the velocity rapidly thermalizes ($\dot{v} \approx 0$), while the spacial coordinate remains far from equilibrium. In the so-called adiabatic transformation, the evolution of the particle's position is thus given by the transformed Langevin equation

$$dx(t) = \frac{\sqrt{2k_B T}}{\gamma_0} dW_t$$

Let us next suppose that we have a Brownian particle in an external field of force (see Figure 3),

generating a potential $U(t, x)$. This leads to the Langevin equation

$$\begin{aligned} \frac{dx(t)}{dt} &= v(t) \\ m dv(t) &= -\gamma_0 v(t) dt - \frac{\partial U}{\partial x}(t, x(t)) + \sqrt{2k_B T \gamma_0} dW_t \end{aligned}$$

In the over-damped limit, after relaxation time, the adiabatic elimination of the fast variables (Gardiner 2004) leads to an equation similar to the one encountered in the previous section:

$$dx(t) = -\frac{1}{\gamma_0} \frac{\partial U}{\partial x}(t, x(t)) + \frac{\sqrt{2k_B T}}{\gamma_0} dW_t$$

In the particular case of some double-well potential $x \rightarrow U(t, x)$ with slow periodic variation, the following patterns of behavior of the solution trajectories will be experienced. If temperature is high, noise has a predominant influence on the motion, and the particle often crosses the barrier separating the two wells during one period. The behavior of the particle does not seem to be periodic but rather chaotic. If temperature is small, the particle stays for a very long time in the starting well, fluctuating weakly around the equilibrium position. It has too low energy to follow the periodic variation of the potential. So in this case too, the trajectories do not look periodic. Between these two extreme situations, there exists a regime of noise intensities for which the energy transmitted by the noise is sufficient to cross the barrier almost twice per period. The parameters are then near to the resonance point and the motion exhibits periodic switching (Figure 4).

Transition Criteria and Quasideterministic Motion

Studying stochastic resonance accordingly means looking for the range of regimes for which periodic behavior is enhanced and eventually optimal. The optimal relation between period T and noise

intensity ε emerges in the small noise limit. To explain this, let us focus on the basic indicator for periodic transitions – the time the Brownian particle needs to exit from the starting well, say the left one. In the “frozen” case, that is, if the time variation of the potential term is eliminated just by freezing it at some time s , the asymptotics of the exit time is derived from the classical large deviation theory of randomly perturbed dynamical systems (see Freidlin and Wentzell 1998). Let us assume that U is locally Lipschitz. We denote by D_l (resp. D_r) the domain corresponding to the left (resp. right) well and χ their common boundary. The law of the first exit time $\tau_{D_l}^\varepsilon = \inf\{t \geq 0, X_t^\varepsilon \notin D_l\}$ is described by some particular functional related to large deviation. For $t > 0$, we introduce the “action functional” on the space of continuous functions $\mathcal{C}([0, t])$ on $[0, t]$ by

$$S_t^\varepsilon(\varphi) = \begin{cases} \frac{1}{2} \int_0^t (\dot{\varphi}_u + \frac{\partial U}{\partial x}(s, \varphi_u))^2 du, & \text{if } \varphi \text{ is abs.} \\ +\infty & \text{continuous} \\ & \text{otherwise} \end{cases}$$

which is non-negative and vanishes on the set of solutions of the ordinary differential equation $\dot{x} = -(\partial U / \partial x)(s, x)$. Let x and $y \in \mathbb{R}$. In relation with the action functional, we define the quasipotential

$$V_s(x, y) = \inf\{S_t^\varepsilon(\varphi) : \varphi \in \mathcal{C}([0, t]), \varphi_0 = x, \varphi_t = y, t \geq 0\}$$

It represents the minimal work the diffusion starting in x has to do in order to reach y . To switch wells, the Brownian particle starting in the left well's bottom x_l has to overcome the barrier. So we let

$$\bar{V}_s = \inf_{y \in \chi} V_s(x_l, y)$$

This minimal work needed to exit from the left well can be computed explicitly, and is seen to equal to twice its depth. The asymptotic behavior of the exit time is expressed by

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \ln \mathbb{E}[\tau_{D_l}^\varepsilon] = \bar{V}_s$$

and

$$\lim_{\varepsilon \rightarrow 0} \mathbb{P}_x \left(e^{(\bar{V}_s - \delta)/\varepsilon} < \tau_{D_l}^\varepsilon < e^{(\bar{V}_s + \delta/\varepsilon)} \right) = 1$$

for any $\delta > 0$

The prefactor for the exponential rate, derived by Freidlin and Wentzell (1998), was first given by Eyring and Kramers and then by Bovier *et al.* (2004).

Let us now assume that the left well is the deeper one at time s . If the Brownian particle has enough time to cross the barrier, that is, if $T > e^{\bar{V}_s/\varepsilon}$, then whatever the starting point is, Freidlin (2000) proved that it should stay near x_l in the following sense:

$$\Lambda(t \in [0, 1] : |X_{tT}^\varepsilon - x_l| > \delta) \rightarrow 0$$

in probability as $\varepsilon \rightarrow 0$. Here Λ denotes Lebesgue measure on \mathbb{R} . If $T < e^{\bar{V}_s/\varepsilon}$, the time left is not long enough for crossings: the particle stays in the starting well, near the stable equilibrium point:

$$\Lambda(t \in [0, 1] : |X_{tT}^\varepsilon - (x_l 1_{\{x \in D_l\}} + x_r 1_{\{x \in D_r\}})| > \delta) \rightarrow 0$$

This observation is at the basis of Freidlin's law of quasideterministic periodic motion discussed in the subsequent section. The lesson it teaches is this: to observe switching of the position to the energetically most favorable well, T should be larger than some critical level $e^{\lambda/\varepsilon}$. Measuring time in exponential scales by μ through the equation $T^\varepsilon = e^{\mu/\varepsilon}$, the condition becomes $\mu > \lambda$.

Stochastic Resonance for Landscapes, Frozen on Half-Periods

This particular case has analytical advantages, since it allows one to employ classical techniques of semigroup and operator theory. The situation is the following: let U be a double-well potential with minima $x_l = -1$ and $x_r = 1$ and a saddle point at the origin. We assume that $U(x) \rightarrow \infty$ as $|x| \rightarrow \infty$ and $U(-1) = -V/2 = -\bar{V}_l/2$, $U(1) = -v/2 = -\bar{V}_r/2$, $U(0) = 0$, and $0 < v < V$. We define the 1-periodic potential by $U(t, x) = U(t + 1/2, -x)$. Hence on each half-period the corresponding diffusion is time homogeneous. The critical level λ is then easily defined by $\lambda = v$, that is, twice the depth of the shallow well. By letting

$$\phi(t) = \begin{cases} -1 & \text{for } t \in [k, k + \frac{1}{2}) \\ 1 & \text{for } t \in [k + \frac{1}{2}, k + 1), \end{cases} \quad k = 0, 1, 2, \dots$$

the periodic function which describes the location of the global minimum of the potential, we get in the small noise limit

$$\Lambda(t \in [0, 1] : |X_{tT}^\varepsilon - \phi(t)| > \delta) \rightarrow 0$$

in probability as $\varepsilon \rightarrow 0$. This result expresses Freidlin's law of quasideterministic motion: for large periods, the trajectories of the particle approach a periodic deterministic function. But the sense in which this notion measures periodicity does not take into account that for large periods short excursions to the wrong well may occur in an erratic way without counting much for Lebesgue measure of time. In fact, if the period is too large, that is, $\mu > V$, the time available in one period permits the exit of not only the shallow well but also that of the deep well. So, whatever the starting position of

the particle is, the number of observed transitions in one half period becomes very large. Indeed the first time ξ the particle starting in x_1 hits again x_1 after visiting the position x_r satisfies

$$E(\xi) = e^{\nu/\varepsilon} + e^{V/\varepsilon} < T^\varepsilon = e^{\mu/\varepsilon}$$

The motion of the particle appears more chaotical than periodic: noise intensity is too large compared to period length. We avoid this range of chaotic spontaneous transitions by defining the resonance interval $I_R = [\nu, V]$, as the range of admissible energy parameters μ for randomly periodic behavior. In this regime, the trajectories possess periodicity properties. In these terms the resonance point describes the tuning rate $\mu_R \in I_R$ for which the stochastic response to weak external periodic forcing is optimal. To make sense, this point has to refer to some measure of quality for periodicity of random trajectories. In the huge physics literature concerning resonance, two families of criteria can be distinguished. The first one is based on invariant measures and spectral properties of the infinitesimal generator associated with the diffusion X^ε . Now, X^ε is not Markovian and consequently does not admit invariant measures. But by taking into account deterministic motion of time in the interval of periodicity and considering the process $Z_t = (t \bmod(T^\varepsilon), X_t)$, we obtain a Markov process with an invariant measure $\nu_t(x)dx$. In other words, the law of $X_t \sim \nu_t(x)dx$ and the law of $X_{t+T} \sim \nu_{t+T}(x)dx$, under this measure, are the same for all $t \geq 0$. Let us present the most important ones:

- the spectral power amplification (SPA) which plays an eminent role in the physics literature describes the energy carried by the spectral component of the averaged trajectories of X^ε corresponding to the period:

$$M_{SPA}(\varepsilon, T) = \left| \int_0^1 E_\nu[X_{sT}^\varepsilon] e^{2\pi i s} ds \right|^2$$

- the SPA-to-noise ratio, giving the ratio of the amplitude of the response and the noise intensity, which is also related to the signal-to-noise ratio:

$$M_{SPN}(\varepsilon, T) = M_{SPA}(\varepsilon, T)/\varepsilon^2$$

- the total energy of the averaged trajectories

$$M_{EN}(\varepsilon, T) = \int_0^1 (E_\nu[X_{sT}^\varepsilon])^2 ds$$

The second family of criteria is more probabilistic. It refers to quality measures based on transition times between the domains of attraction of the local minima, residence times distributions measuring the time spent in one well between two transitions, or interspike times. This family is certainly less popular in the physics community.

However, measures related to invariant measures may suffer from robustness deficiency (Imkeller and Pavlyukevich 2002). To explain what we mean by robustness, let us introduce a model reduction first discussed by McNamara and Wiesenfeld (1989). Instead of studying the diffusion X^ε in the double-well landscape, they introduce a two-state Markov chain Y^ε (Figure 5) the dynamics of which just takes account of the domain of attraction the diffusion is in, and therefore with state space $\{-1, 1\}$. A reasonable choice of the infinitesimal generator should retain the dynamics of the diffusion's transitions characterized by Kramers' rate. We may take

$$Q(t) = \begin{pmatrix} -\varphi & \varphi \\ \psi & -\psi \end{pmatrix}, \quad 0 \leq t \leq \frac{T}{2}$$

$$Q(t) = \begin{pmatrix} -\psi & \psi \\ \varphi & -\varphi \end{pmatrix}, \quad \frac{T}{2} \leq t < T$$

periodically continued on \mathbb{R}_+ . Here, $\varphi = pe^{-V/\varepsilon}$ and $\psi = qe^{-\nu/\varepsilon}$. The prefactors of subexponential order are beyond the scope of large deviation theory. They are related to the curvature of the potential in the

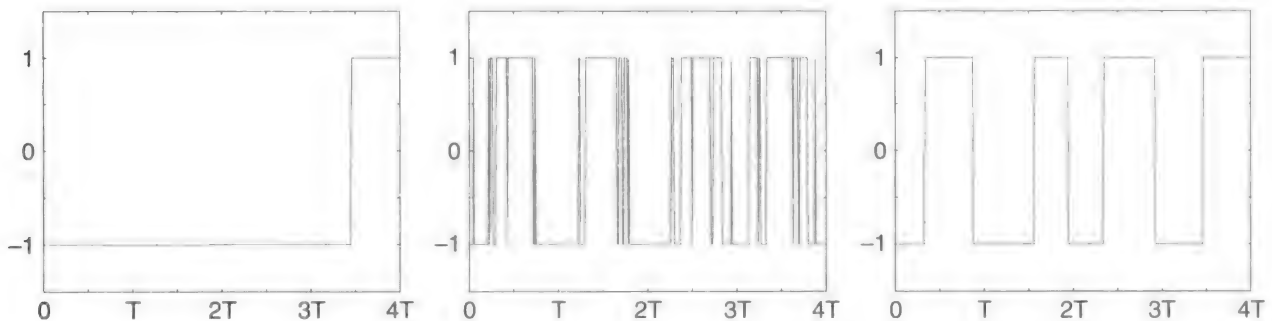


Figure 5 Resonance pictures for Markov chain.

minima and the saddle point of the landscape and given by

$$p = \frac{1}{2\pi} \sqrt{U'''(-1)|U''(0)|}$$

$$q = \frac{1}{2\pi} \sqrt{U'''(1)|U''(0)|}$$

On the intervals $[kT/2, (k+1)T/2[, k \geq 0$, the Markov chain Y^ε is time-homogeneous and its transition probabilities can be expressed in terms of φ and ψ . For instance, the probability with which the chain jumps from state -1 to state $+1$ in the time window $[t, t+h]$ equals $\varphi h + o(h)$, if this time interval is contained in $[kT/2, (k+1)T/2[$ for some even k . The stationary measure of the Markov chain denoted by ν can be explicitly calculated, and so can the classical quality measures based on the spectral notions. For instance, the spectral power amplification coefficient equals

$$M_{\text{SPA}}(\varepsilon, T) = \left| \int_0^1 \mathbb{E}_\nu[Y_{st}^\varepsilon] e^{2\pi i s} ds \right|^2$$

$$= \frac{4}{\pi^2} \frac{T^2(\varphi - \psi)^2}{(\varphi + \psi)^2 T^2 + \pi^2}$$

This simple expression admits asymptotically a unique maximum which exhibits the resonance point:

$$T_{\text{opt}}^\varepsilon = \frac{\pi}{\sqrt{2pq}} \sqrt{\frac{\nu}{V-\nu}} e^{(V+\nu)/2\varepsilon} \left\{ 1 + \mathcal{O}(e^{-(V-\nu)/\varepsilon}) \right\}$$

The optimal period is then exponentially large – as was suggested by large deviation theory – and the growth rate is the sum of the two wells' depths. The simple Markov chain model is popular since the usual physical quantities are easy computable and since it is believed to mimic the dynamics of a Brownian particle in the corresponding double-well landscape. However, the models are not as similar as expected (Freidlin 2003). Indeed, in a reasonably large time window around the resonance point for Y^ε , the tuning picture of the spectral power amplification for the diffusion is different. Under weak regularity conditions on the potential, it exhibits strict monotonicity in the window. Hence, optimal tuning points for diffusion and Markov chain differ essentially. In other words, the SPA tuning behavior of the diffusion is not robust for passage to the reduced model. This strange deficiency is difficult to explain. The main reason of this subtle effect appears to be that the diffusive nature of the Brownian particle is neglected in the reduced model. In order to point out this feature, we may compute the SPA coefficient of $g(X^\varepsilon)$, where g is a

particular function designed to cut out the small fluctuations of the diffusion in the neighborhood of the bottoms of the wells, by identifying all states there. So $g(x) = -1$ (resp. 1) in some neighborhood of -1 (resp. 1) and otherwise g is the identity. This results in

$$\tilde{M}_{\text{SPA}}(\varepsilon, T) = \left| \int_0^1 \mathbb{E}_\nu g(X_{sT}^\varepsilon) e^{2\pi i s} ds \right|^2$$

In the small noise limit this quality function admits a local maximum close to the resonance point of the reduced model: the growth rate of $T_{\text{opt}}^\varepsilon$ is also given by the sum of the wells' depths. So the lack of robustness seems to be due to the small fluctuations of the particle in the wells' bottoms. In any case, this clearly calls for other quality measures to be used to transfer properties of the reduced model to the original one. Our discussion indicates that due to their emphasis on the pure transition dynamics, the second family of quality measures should be used. For these notions there is no need to restrict to landscapes frozen in time-independent potential states on half-period intervals.

Stochastic Resonance for Continuously Varying Landscapes

From now on the potential $U(t, x)$ is supposed to be continuously varying in (t, x) . For simplicity, its local minima are assumed to be located at ± 1 , and its only saddle point at 0 , independently of time. So the only meta-stable states on the whole time axis are ± 1 . Let us denote by $\Delta_-(t)$ (resp. $\Delta_+(t)$) the depth of the left (resp. right) well at time t . Together with U , these functions are continuous and 1-periodic. Assume that they are strictly monotonous between their global extrema. Let us now come back to the motion of a Brownian particle in this landscape. The exit time law by Eyring–Kramers–Freidlin entails that trajectories get close to the global minimum, if the period is large enough. Stated as before in exponential rates $T = e^{\mu/\varepsilon}$, with $\mu \geq \max_{i=\pm} \sup_{t \geq 0} 2\Delta_\pm(t)$, that is, μ exceeds the maximal work needed to cross the barrier, the particle often switches between the two wells and should stay close to the deepest position in the landscape. This position being described by the function $\phi(t) = 21_{\{\Delta_+(t) > \Delta_-(t)\}} - 1$, we get in the small noise limit

$$\Lambda(t \in [0, 1] : |X_{tT}^\varepsilon - \phi(t)| > \delta) \rightarrow 0$$

in probability. But on these long timescales, many short excursions to the wrong well are observed, and trajectories look chaotic instead of periodic. So we

have to look at smaller periods even at the cost that the particle may not stay close to the global minimum. Let us study the transition dynamics. Assume that the starting point is -1 corresponding to the bottom of the deep well. If the depth of the well is always larger than $\mu = \varepsilon \log T^\varepsilon$, the particle has too little time during one period to climb the barrier, and should stay in the starting well. If, on the contrary, the minimal work to leave the starting well, given by $2\Delta_-(s)$, becomes smaller than μ at some time s , then the transition can and will happen. More formally, for $\mu \in [\inf_{t \geq 0} 2\Delta_-(t), \sup_{t \geq 0} 2\Delta_-(t)]$, we define (Figure 6).

$$a_\mu^-(s) = \inf\{t \geq s : 2\Delta_-(t) \leq \mu\}$$

The first transition time from -1 to 1 denoted τ_+ has the following asymptotic behavior as $\varepsilon \rightarrow 0$: $\tau_+/T^\varepsilon \rightarrow a_\mu^-(0)$. At the second transition the particle returns to the starting well. If a_μ^+ is defined analogously with respect to the depth function Δ_+ , this transition will occur near the deterministic time $a_\mu^+(a_\mu^-(s))T^\varepsilon$. In order to observe periodicity, and to exclude chaoticity from all parts of its trajectories, the particle has to stay for some time in the other well before returning. This will happen under the assumption $2\Delta_+(a_\mu(0)) > \mu$, that is, the right well is the deep one at transition time. In fact, we can define the resonance interval I_R (Figure 7), as the set of all scales μ for which trajectories exhibit periodicity in the small noise limit, by

$$I_R = \left[\max_{i=\pm} \inf_{t \geq 0} 2\Delta_i(t), \inf_{t \geq 0} \max_{i=\pm} 2\Delta_i(t) \right]$$

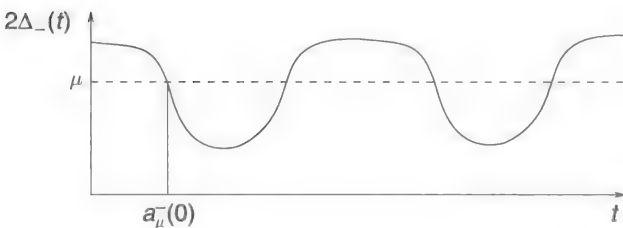


Figure 6 Definition of a_μ^- .

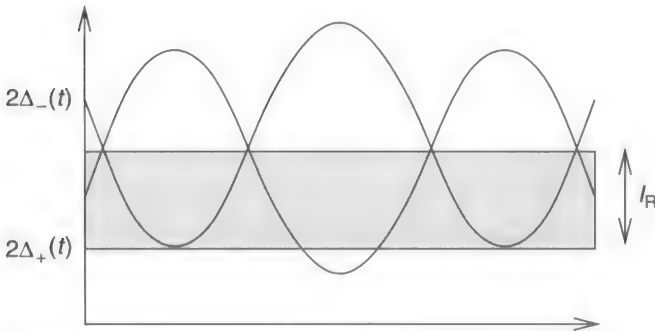


Figure 7 Resonance interval.

On this interval they get close to deterministic periodic ones. Again, periodicity is quantified by a quality measure, to be maximized in order to obtain resonance as the best possible response to periodic forcing. One interesting measure is based on the probability that random transitions happen in some small time window around a deterministic time, in the small noise limit (Herrmann and Inkeller 2005). Formally, for $h > 0$, the measure gives

$$M_b(\varepsilon, T) = \min_{i=\pm} \mathbb{P}_i(\tau_\mp/T^\varepsilon \in [a_\mu^i - h, a_\mu^i + h])$$

where \mathbb{P}_i is the law of the diffusion starting in i . In the small noise limit, this quality measure tends to 1, and optimal tuning can be related to the exponential rate at which this happens. This is due to the following large deviations principle:

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log(1 - M_b(\varepsilon, T)) = \max_{i=\pm} \{\mu - 2\Delta_i(a_\mu^i - h)\}$$

for $\mu \in I_R$, with uniform convergence on each compact subset of I_R . The result is established using classical large-deviation techniques applied to locally time homogeneous approximations of the diffusion. Maximizing the transition probability in the time window position means minimizing the default rate obtained by the large deviations principle. This can be easily achieved. In fact, if the window length $2h$ is small, then $\mu - 2\Delta_i(a_\mu^i - h) \approx 2h\Delta'_i(a_\mu^i)$, since $2\Delta_i(a_\mu^i) = \mu$ by definition. The value $\Delta'_i(a_\mu^i)$ is negative, so we have to find the position where its absolute value is maximal. In this position the depth of the starting well has the most rapid drop under the level μ , characterizing the link between the noise intensity and the period. So the transition time is best concentrated around it.

It is clear that a good candidate for the resonance point is given by the eventually existing limit of the global minimizer $\mu_R(h)$ as the window length h tends to 0. This limit is therefore called the resonance point of the diffusion with time-periodic landscape U . Let us note that for sinusoidal depth functions

$$\Delta_-(t) = \frac{V+v}{4} + \frac{V-v}{4} \cos(2\pi t)$$

and

$$\Delta_+(t) = \Delta_-(t + \pi)$$

the optimal tuning is given by $T^\varepsilon = \exp \mu_R/\varepsilon$ with $\mu_R = (v + V)/2$. This optimal rate is equivalent to the optimal rate given by the SPA coefficient of the reduced dynamics' Markov chain in the preceding section.

The big advantage of the quality measure M_b is its robustness. Indeed, consider the reduced model

consisting of a two-state Markov chain with infinitesimal generator

$$Q(t) = \begin{pmatrix} -\varphi(t) & \varphi(t) \\ \psi(t) & -\psi(t) \end{pmatrix}$$

where $\varphi(t) = \exp -2\Delta_-(t/T)/\varepsilon$ and $\psi(t) = \exp -2\Delta_+(t/T)/\varepsilon$. The law of transition times of this Markov chain is readily computed from Laplace transforms. Normalized by T^ε it converges to a_μ^i . This calculation even reveals a rigorous underlying pattern for the second- and higher-order transition times interpreting the interspike distributions of the physics literature. The dynamics of diffusion and Markov chain are similar. Resonance points provided by M_b for the diffusion and its analog for the Markov chain agree.

Related Notions: Synchronization

In the preceding sections, we interpreted stochastic resonance as optimal response of a randomly perturbed dynamical system to weak periodic forcing, in the spirit of the physics literature (see Gammaitoni *et al.* (1998)). Our crucial assumption concerned the barrier heights a Brownian particle has to overcome in the potential landscape of the dynamical system: it is uniformly lower bounded in time. Measures for the quality of tuning were based on essentially two concepts: one concerning spectral criteria, with the spectral power amplification as most prominent member, the other one concerning the pure transitions dynamics between the domains of attraction of the local minima. A number of different criteria can be used to create an optimal tuning between the intensity of the noise perturbation and the large period of the dynamical system. The relations have to be of an exponential type $T = \exp \mu/\varepsilon$, since the Brownian particle needs exponentially long times to cross the barrier separating the wells according to the Eyring–Kramers–Freidlin transition law. Our barrier height assumption seems natural in many situations, but can fail in others. If it becomes small periodically, and eventually scales with the noise-intensity parameter, the Brownian particle does not need to wait an exponentially long time to climb it. So periodicity obtains for essentially smaller timescales. In this setting, the slowness of periodic forcing may also be assumed to be essentially subexponential in the noise intensity.

If it is fast enough to allow for substantial changes before large deviation effects can take over, we are in the situation of Berglund and Gentz (2002). They in fact consider the case in which the barrier between the wells becomes low twice per period,

to the effect of modulating periodically a bifurcation parameter: at time zero the right-hand well becomes almost flat, and at the same time the bottom of the well and the saddle approach each other; half a period later, a spatially symmetric scenario is encountered. In this situation, there is a threshold value for the noise intensity under which transitions become unlikely. Above this threshold, the trajectories typically contain two transitions per period. Results are formulated in terms of concentration properties for random trajectories. The intuitive picture is this: with overwhelming probability, sample paths will be concentrated in spacetime sets scaling with the small parameters of the problem. In higher dimensions, these sets may be given by adiabatic or center manifolds of the deterministic system, which allow model reduction of higher-dimensional systems to lower-dimensional ones. Asymptotic results hold for any choice of the small parameters in a whole parameter region. A passage to the small noise limit as for optimal tuning in the preceding sections is not needed.

Related problems studied by Berglund and Gentz in the multidimensional case concern the noise-induced passage through periodic orbits, where unexpected phenomena arise. Here, as opposed to the classical Freidlin–Wentzell theory, the distribution of first-exit points depends nontrivially on the noise intensity. Again aiming at results valid for small but nonvanishing parameters in subexponential scale ranges, they investigate the density of first-passage times in a large regime of parameter values, and obtain insight into the transition from the stochastic resonance regime into the synchronization regime.

See also: Dynamical Systems in Mathematical Physics: An Illustration from Water Waves; Magnetic Resonance Imaging; Spectral Theory for Linear Operators; Stochastic Differential Equations.

Further Reading

- Benzi R, Sutera A, and Vulpiani A (1981) The mechanism of stochastic resonance. *Journal of Physics A* 14: L453–L457.
- Berglund N and Gentz B (2002) A sample-paths approach to noise-induced synchronization: stochastic resonance in a double-well potential. *Annals of Applied Probability* 12(4): 1419–1470.
- Bovier A, Eckhoff M, Gaynard V, and Klein M (2004) Metastability in reversible diffusion processes. I. Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society* 6(4): 399–424.
- Freidlin MI (2000) Quasi-deterministic approximation, metastability and stochastic resonance. *Physica D* 137(3–4): 333–352.

- Freidlin M (2003) Noise sensitivity of stochastic resonance and other problems related to large deviations. In: Sri Namachchivaya and Lin YK (eds.) *IUTAM Symposium on Nonlinear Stochastic Dynamics*, Solid Mech. Appl, vol. 110, pp. 43–55. Dordrecht: Kluwer Academic.
- Freidlin MI and Wentzell AD (1998) *Random Perturbations of Dynamical Systems*, 2nd edn. New York: Springer.
- Gammaitoni L, Hänggi P, Jung P, and Marchesoni F (1998) Stochastic resonance. *Reviews of Modern Physics* 70(1): 223–287.
- Gardiner CW (2004) *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Springer Series in Synergetics, 3rd edn., vol. 13. Berlin: Springer.
- Herrmann S and Imkeller P (2005) The exit problem for diffusions with time-periodic drift and stochastic resonance. *Annals of Applied Probability* 15(1A): 39–68.
- Imkeller P and Pavlyukevich I (2002) Model reduction and stochastic resonance. *Stochastics and Dynamics* 2(4): 463–506.
- Mazo RM (2002) *Brownian Motion: Fluctuations, Dynamics, and Applications*. International Series of Monographs on Physics, vol. 112. New York: Oxford University Press.
- McNamara B and Wiesenfeld K (1989) Theory of stochastic resonance. *Physical Review A* 39(9): 4854–4869.
- Nelson E (1967) *Dynamical Theories of Brownian Motion*. Princeton, NJ: Princeton University Press.
- Schweitzer F (2003) *Brownian Agents and Active Particles*, Springer Series in Synergetics, (Collective dynamics in the natural and social sciences, With a foreword by J. Dooyne Farmer). Berlin: Springer.

Strange Attractors see Lyapunov Exponents and Strange Attractors

String Field Theory

L Rastelli, Princeton University, Princeton, NJ, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

String field theory (SFT) is the second-quantized approach to string theory. In the usual, first-quantized, formulation of string perturbation theory, one postulates a recipe for the string S -matrix in terms of a sum over two-dimensional (2D) world sheets embedded in spacetime. Very schematically,

$$\langle\langle V_1(k_1) \dots V_n(k_n) \rangle\rangle = \sum_{\text{topologies}} g_s^{-\chi} \int [d\mu_\alpha] \langle V_1(k_1) \dots V_n(k_n) \rangle_{\{\mu_\alpha\}} \quad [1]$$

Here the left-hand side stands for the S -matrix of the physical string states $\{V_a(k_a)\}$. The symbol $\langle \dots \rangle_{\{\mu_\alpha\}}$ denotes a correlation function on the 2D world sheet, which is a punctured Riemann surface of Euler number χ and given moduli $\{\mu_\alpha\}$. In SFT, one aims to recover this standard prescription from the Feynman rules of a second-quantized spacetime action $S[\Phi]$. The string field Φ , the fundamental dynamical variable, can be thought of as an infinite-dimensional array of spacetime fields $\{\phi^i(x^\mu)\}$, one field for each basis state in the Fock space of the first-quantized string.

The most straightforward way to construct $S[\Phi]$ uses the unitary light-cone gauge. Light-cone SFT is an almost immediate transcription of Mandelstam's

light-cone diagrams in a second-quantized language. While often useful as a bookkeeping device, light-cone SFT seems unlikely to represent a real improvement over the first-quantized approach. By contrast, from our experience in ordinary quantum field theory, we should expect Poincaré-covariant SFTs to give important insights into the issues of vacuum selection, background independence and the nonperturbative definition of string theory.

Covariant SFT actions are well established for the open (Witten 1986), closed (Zwiebach 1993) and open/closed (Zwiebach 1998) bosonic string. These theories are based on the BRST formalism, where the world sheet variables include the bc ghosts introduced in gauge-fixing the world sheet metric to the conformal gauge $g_{ab} \sim \delta_{ab}$. (An alternative approach (Hata *et al.*), based on covariantizing light-cone SFT, will not be described in this article.) Much less is presently known for the superstring: classical actions have been established for the Neveu-Schwarz sector of the open superstring (Berkovits 2001) and for the heterotic string (Berkovits *et al.* 2004).

During the first period of intense activity in SFT (1985–1992), the covariant bosonic actions were constructed and shown to pass the basic test of reproducing the S -matrix [1] to each order in the perturbative expansion. The more recent revival of the subject (since 1999) was triggered by the realization that SFT contains nonperturbative information as well: D-branes emerge as solitonic solutions of the classical equations of motion in

open SFT (OSFT). We can hope that the nonperturbative string dualities will also be understood in the framework of SFT, once covariant SFTs for the superstring are better developed.

In this article, we review the basic formalism of covariant SFT, using for illustration purposes the simplest model – cubic bosonic OSFT. We then briefly sketch the generalization to bosonic SFTs that include closed strings. Finally, we turn to the subjects of classical solutions in OSFT and the physics of the open-string tachyon.

Open Bosonic SFT

The standard formulation of string theory starts with the choice of an on-shell spacetime background where strings propagate. In the bosonic string, the closed string background is described by a conformal field theory of central charge 26 (the “matter” CFT). The total world sheet CFT is the direct sum of this matter CFT and of the universal ghost CFT, of central charge -26 . To describe open strings, we must further specify boundary conditions for the string endpoints. The open-string background is encoded in a boundary CFT (BCFT), a CFT defined in the upper-half plane, with conformal boundary conditions on the real axis (see Boundary Conformal Field Theory in this encyclopedia). In modern language, the choice of BCFT corresponds to specifying a D-brane state.

In classical OSFT, we fix the closed-string background (the bulk CFT) and consider varying the D-brane configuration (the boundary conditions). To lowest order in g_s , we can neglect the back-reaction of the D-brane on the closed-string fields, since this is a quantum effect from the open-string viewpoint. Let us prepare the ground by recalling the standard σ -model philosophy. To describe off-shell open-string configurations, we should allow for general (not necessarily conformal) boundary conditions. We can imagine to proceed as follows:

1. We choose an initial open-string background, a reference BCFT that we shall call BCFT_0 . For example, a Dp brane in flat 26 dimensions (Neumann boundary conditions on $p+1$ coordinates, Dirichlet on $25-p$ coordinates).
2. We then write a basis of boundary perturbations around this background. Taking, for example, BCFT_0 to be a D25 brane in flat space, the world sheet action S_{WS} takes the schematic form

$$S_{\text{WS}} = \frac{1}{2\pi\alpha'} \int_{\text{UHP}} \partial X_\mu \bar{\partial} X^\mu + \int_R \tilde{T}(x^\mu) + \tilde{A}_\nu(x^\mu) \partial X^\nu + \tilde{B}_\nu(x^\mu) \partial^2 X^\nu + \dots \quad [2]$$

Here to the standard free bulk action (integrated over the upper-half complex plane UHP) we have added perturbation localized on the real axis R . Notice that the basis of perturbations depends on the chosen BCFT_0 .

3. We interpret the coefficients $\{\tilde{\phi}^i(x^\mu)\}$ of the perturbations as spacetime fields. (The tilde on $\tilde{\phi}^i(x)$ serves as a reminder that these fields are not quite the same as the fields $\phi^i(x)$ that will appear in the OSFT action). We are after a spacetime action $S[\{\tilde{\phi}^i\}]$ such that solutions of its classical equations of motion correspond to conformal boundary conditions:

$$\begin{aligned} \frac{\delta S}{\delta \tilde{\phi}^i} &= 0(\text{spacetime}) \\ \Leftrightarrow \beta_i[\{\tilde{\phi}^j\}] &= 0 \text{ (world sheet)} \end{aligned} \quad [3]$$

We recognize in [2] the familiar open-string tachyon $\tilde{T}(x)$ and gauge field $\tilde{A}_\mu(x)$, which are the lowest modes in an infinite tower of fields. Relevant perturbations on the world sheet (with conformal dimension $h < 1$) correspond to tachyonic fields in spacetime ($m^2 < 0$), whereas marginal world sheet perturbations ($h = 1$) give massless spacetime fields. To achieve a complete description, we must include all the higher massive open-string modes as well, which correspond to nonrenormalizable boundary perturbations ($h > 1$). In the traditional σ -model approach, this appears like a daunting task. The formalism of OSFT will automatically circumvent this difficulty.

The Open-String Field

In covariant SFT the reparametrization ghosts play a crucial role. The ghost CFT consists of the Grassmann odd fields $b(z), c(z), \bar{b}(\bar{z}), \bar{c}(\bar{z})$, of dimensions $(2, 0), (-1, 0), (0, 2), (0, -1)$, respectively. The boundary conditions on the real axis are $b = \bar{b}, c = \bar{c}$. The state space $\mathcal{H}_{\text{BCFT}_0}$ of the full matter+ghost BCFT can be broken up into subspaces of definite ghost number,

$$\mathcal{H}_{\text{BCFT}_0} = \bigoplus_{G=-\infty}^{\infty} \mathcal{H}_{\text{BCFT}_0}^{(G)} \quad [4]$$

We use conventions where the $\text{SL}(2, \mathbf{R})$ vacuum $|0\rangle$ carries zero ghost number, $G(|0\rangle) = 0$, while $G(c) = +1$ and $G(b) = -1$. As is familiar from the first-quantized treatment, physical open-string states are identified with $G = +1$ cohomology classes of the BRST operator,

$$\begin{aligned} Q|V_{\text{phys}}\rangle &= 0, & |V_{\text{phys}}\rangle &\sim |V_{\text{phys}}\rangle + Q|\Lambda\rangle \\ G(|V_{\text{phys}}\rangle) &= +1 \end{aligned} \quad [5]$$

where the nilpotent BRST operator Q has the standard expression

$$Q = \frac{1}{2\pi i} \oint (c T_{\text{matter}} + :bc\partial c:) \quad [6]$$

Though not *a priori* obvious, it turns out that the simplest form of the OSFT action is achieved by taking as the fundamental off-shell variable an arbitrary $G=+1$ element of the first-quantized Fock space,

$$|\Phi\rangle \in \mathcal{H}_{\text{BCFT}_0}^{(1)} \quad [7]$$

By the usual state-operator correspondence of CFT, we can also represent $|\Phi\rangle$ as a local (boundary) vertex operator acting on the vacuum,

$$|\Phi\rangle = \Phi(0)|0\rangle \quad [8]$$

The open-string field $|\Phi\rangle$ is really an infinite-dimensional array of spacetime fields. We can make this transparent by expanding it as

$$|\Phi\rangle = \sum_i \int d^{p+1}k |\Phi_i(k)\rangle \phi^i(k_\mu) \quad [9]$$

where $\{|\Phi_i(k)\rangle\}$ is some convenient basis of $\mathcal{H}_{\text{BCFT}_0}^{(1)}$ that diagonalizes the momentum k_μ . The fields ϕ^i are *a priori* complex. This is remedied by imposing a suitable reality condition on the string field, which will be stated momentarily. Notice that there are many more elements in $\{|\Phi_i(k)\rangle\}$ than in the physical subspace (the cohomology classes of Q). Some of the extra fields will turn out to be nondynamical and could be integrated out, but at the price of making the OSFT action look much more complicated.

It is often useful to think of the string field in terms of its Schrödinger representation, that is, as a functional on the configuration space of open strings. Consider the unit half-disk in the upper-half plane, $D_H \equiv \{|z| \leq 1, \Im z \geq 0\}$, with the vertex operator $\Phi(0)$ inserted at the origin. Impose BCFT_0 open string boundary conditions for the fields $X(z, \bar{z})$ on the real axis (here $X(z, \bar{z})$ is a short-hand notation for all matter and ghost fields), and boundary conditions $X(\sigma) = X_b(\sigma)$ on the curved boundary of D_H , $z = \exp(i\sigma)$, $0 \leq \sigma \leq \pi$. The path integral over $X(z, \bar{z})$ in the interior of the half-disk assigns a complex number to any given $X_b(\sigma)$, so we obtain a functional $\Phi[X_b(\sigma)]$. This is the Schrödinger wave function of the state $\Phi(0)|0\rangle$. Thus, we can think of open-string functionals $\Phi[X_b(\sigma)]$ as the fundamental variables of OSFT. This is as it should be: the

first-quantized wave functions are promoted to dynamical fields in the second-quantized theory. Finally, let us quote the reality condition for the string field, which takes a compact form in the Schrödinger representation:

$$\begin{aligned} \Phi[X^\mu(\sigma), b(\sigma), c(\sigma)] \\ = \Phi^*[X^\mu(\pi - \sigma), b(\pi - \sigma), c(\pi - \sigma)] \end{aligned} \quad [10]$$

where the superscript $*$ denotes complex conjugation.

The Classical Action

With all the ingredients in place, it is immediate to write the quadratic part of the OSFT action. The linearized equations of motion must reproduce the physical-state condition [5]. This suggests

$$S \sim \langle \Phi | Q | \Phi \rangle \quad [11]$$

Here $\langle | \rangle$ is the usual BPZ inner product of BCFT_0 , which is defined in terms of a two-point correlator on the disk, as we review below. The ghost anomaly implies that on the disk we must have $G_{\text{tot}} = +3$, which happily is the case in [11]. Moreover, since the inner product is nondegenerate, variation of [11] gives

$$Q|\Phi\rangle = 0 \quad [12]$$

as desired. The equivalence relation $|V_{\text{phys}}\rangle \sim |V_{\text{phys}}\rangle + Q|\Lambda\rangle$ is interpreted in the second-quantized language as the spacetime gauge invariance

$$\delta_\Lambda |\Phi\rangle = Q|\Lambda\rangle, \quad |\Lambda\rangle \in \mathcal{H}_{\text{BCFT}_0}^{(0)} \quad [13]$$

valid for the general off-shell field. This equation is a very compact generalization of the linearized gauge invariance for the massless gauge field. Indeed, focusing on the level-zero components, $|\Phi\rangle \sim A_\mu(x)(c\partial X^\mu)(0)|0\rangle$ and $|\Lambda\rangle \sim \lambda(x)|0\rangle$, we find $\delta A_\mu(x) = \partial_\mu \lambda(x)$. It is then plausible to guess that the nonlinear gauge invariance should take the form

$$\delta_\Lambda |\Phi\rangle = Q|\Lambda\rangle + |\Phi\rangle * |\Lambda\rangle - |\Lambda\rangle * |\Phi\rangle \quad [14]$$

where $*$ is some suitable product operation that conserves ghost number

$$*: \mathcal{H}_{\text{BCFT}_0}^{(n)} \otimes \mathcal{H}_{\text{BCFT}_0}^{(m)} \rightarrow \mathcal{H}_{\text{BCFT}_0}^{(n+m)} \quad [15]$$

Based on a formal analogy with 3D nonabelian Chern–Simons theory, Witten proposed the cubic action

$$S = -\frac{1}{g_o^2} \left(\frac{1}{2} \langle \Phi | Q | \Phi \rangle + \frac{1}{3} \langle \Phi | \Phi * \Phi \rangle \right) \quad [16]$$

The string field $|\Phi\rangle$ is analogous to the Chern–Simons gauge potential $A = A_i dx^i$, the $*$ product to the \wedge product of differential forms, Q to the exterior derivative d , and the ghost number G to the degree

of the form. The analogy also suggests a number of algebraic identities:

$$\begin{aligned}
 Q^2 &= 0 \\
 \langle QA|B \rangle &= -(-1)^{G(A)} \langle A|QB \rangle \\
 Q(A * B) &= (QA) * B + (-1)^{G(A)} A * (QB) \\
 \langle A|B \rangle &= (-1)^{G(A)G(B)} \langle B|A \rangle \\
 \langle A|B * C \rangle &= \langle B|C * A \rangle \\
 A * (B * C) &= (A * B) * C
 \end{aligned} \quad [17]$$

Note in particular the associativity of the $*$ -product. It is straightforward to check that this algebraic structure implies the gauge invariance of the cubic action under [14]. A $*$ -product satisfying all required formal properties can indeed be defined. The most intuitive presentation is in the functional language. Given an open-string curve $X(\sigma)$, $0 \leq \sigma \leq \pi$, we single out the string mid-point $\sigma = \pi/2$ and define the left and right “half-string” curves

$$\begin{aligned}
 X_L(\sigma) &\equiv X(\sigma) & \text{for } 0 \leq \sigma \leq \frac{\pi}{2} \\
 X_R(\sigma) &\equiv X(\pi - \sigma) & \text{for } \frac{\pi}{2} \leq \sigma \leq \pi
 \end{aligned} \quad [18]$$

A functional $\Phi[X(\sigma)]$ can, of course, be regarded as a functional of the two half-strings, $\Phi[X] \rightarrow \Phi[X_L, X_R]$. We define

$$(\Phi_1 * \Phi_2)[X_L, X_R] \equiv \int [dY] \Phi_1[X_L, Y] \Phi_2[Y, X_R] \quad [19]$$

where $\int [dY]$ is meant as the functional integral over the space of half-strings $Y(\sigma)$, with $Y(\pi/2) = X_L(\pi/2) = X_R(\pi/2)$. **Figure 1a** shows two open strings interacting (to form a single open string) if and only if the right half of the first string precisely overlaps with the left half of the second string. Associativity is transparent (**Figure 1b**).

We can now translate this formal construction in the precise CFT language. Very generally, an n -point vertex of open strings can be defined by specifying an n -punctured disk, that is, a disk with marked points on the boundary (punctures) and a choice of local coordinates around each puncture. Local

coordinates are essential since we are dealing with off-shell open-string states. The BPZ inner product (two-point vertex) is given by

$$\begin{aligned}
 \langle \Phi_1 | \Phi_2 \rangle &\equiv \langle I \circ \Phi_1(0) \Phi_2(0) \rangle_{\text{UPH}} \\
 I(z) &= -\frac{1}{z}
 \end{aligned} \quad [20]$$

The symbol $f \circ \Phi(0)$, where f is a complex map, means the conformal transform of $\Phi(0)$ by f . For example, if Φ is a dimension- d primary field, then $f \circ \Phi(0) = f'(0)^d \Phi(f(0))$. If Φ is nonprimary, the transformation rule will be more complicated and involve extra terms with higher derivatives of f . By performing the $\text{SL}(2, \mathbb{C})$ transformation

$$w = h(z) \equiv \frac{1 + iz}{1 - iz} \quad [21]$$

we can represent the two-point vertex as a correlator on the unit disk $D = \{|w| \leq 1\}$,

$$\begin{aligned}
 \langle \Phi_1 | \Phi_2 \rangle &= \langle f_1 \circ \Phi_1(0), f_2 \circ \Phi_2(0) \rangle_D \\
 f_1(z_1) &= -h(z_1), \quad f_2(z_2) = h(z_2)
 \end{aligned} \quad [22]$$

The vertex operators are inserted as $w = -1$ and $w = +1$ on D (see **Figure 2a**) and correspond to the two open strings at (Euclidean) world sheet time $\tau = -\infty$ (we take $z = \exp(i\sigma + \tau)$). The left half of D is the world sheet of the first open string; the right half of D is the world sheet of the second string. The two strings meet at $\tau = 0$ on the imaginary w axis. The three-point Witten vertex is given by

$$\begin{aligned}
 \langle \Phi_1, \Phi_2, \Phi_3 \rangle &\equiv \langle g_1 \circ \Phi_1(0) g_2 \circ \Phi_2(0) g_3 \circ \Phi_3(0) \rangle_D \\
 &\equiv \langle g_1 \circ \Phi_1(0) g_2 \circ \Phi_2(0) g_3 \circ \Phi_3(0) \rangle_D
 \end{aligned} \quad [23]$$

where

$$\begin{aligned}
 g_1(z_1) &= e^{2\pi i/3} \left(\frac{1 + iz_1}{1 - iz_1} \right)^{2/3} \\
 g_2(z_2) &= \left(\frac{1 + iz_2}{1 - iz_2} \right)^{2/3} \\
 g_3(z_3) &= e^{-2\pi i/3} \left(\frac{1 + iz_3}{1 - iz_3} \right)^{2/3}
 \end{aligned} \quad [24]$$

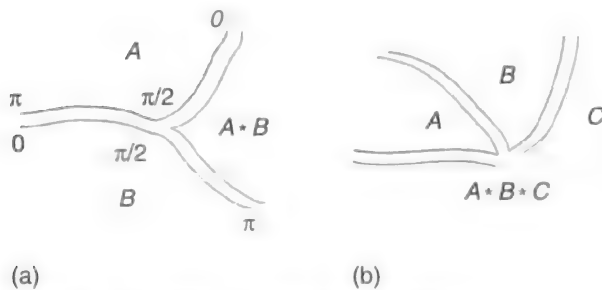


Figure 1 Midpoint overlaps of open strings.

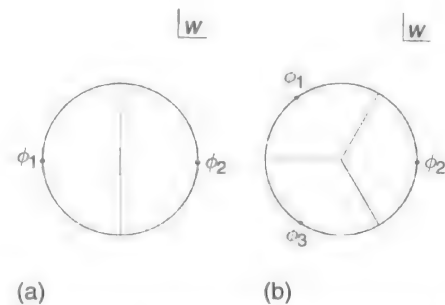


Figure 2 Representation of the quadratic and cubic vertices as 2- and 3-punctured unit disks.

The 3-punctured disk is depicted in Figure 2b, and describes the symmetric mid-point overlap of the three strings at $\tau=0$. Finally, the relation between the three-point vertex and the $*$ -product is

$$\langle \Phi_1 | \Phi_2 * \Phi_3 \rangle \equiv \langle \Phi_1, \Phi_2, \Phi_3 \rangle \quad [25]$$

Knowledge of the right-hand side (RHS) in [25] for all Φ allows to reconstruct the $*$ -product. All formal properties [17] are easily shown to hold in the CFT language. This completes the definition of the OSFT action.

Evaluation of the classical action is completely algorithmic and can be carried out for arbitrary massive states, with no fear of divergences, since in all required correlators the operators are inserted well apart from each other.

Quantization

Quantization is defined by the path integral over the second-quantized string field. The first step is to deal with the gauge invariance [14] of the classical action. The gauge symmetry is reducible: not all gauge parameters $\Lambda^{(0)}$ (the superscript labels ghost number) lead to a gauge transformation. This is clear at the linearized level; indeed, if $\Lambda^{(0)} = Q\Lambda^{(-1)}$, then $\delta_{\Lambda^{(0)}}\Phi^{(1)} = Q^2\Lambda^{(0)} = 0$. Thus, the set $\{\Lambda^{(0)}\}$ gives a redundant parametrization of the gauge group. Characterizing this redundancy is somewhat subtle, since fields of the form $\Lambda^{(-1)} = Q\Lambda^{(-2)}$ do not really lead to a redundancy in $\Lambda^{(0)}$, and so on, *ad infinitum*. It is clear that we need to introduce an infinite tower of (second-quantized) ghosts for ghosts.

The Batalin–Vilkovisky formalism is a powerful way to handle the problem. The basic object is the master action $S(\phi^s, \phi_s^*)$, which is a function of the “fields” ϕ^s and of the “antifields” ϕ_s^* . Each field is paired with a corresponding antifield of opposite Grassmanality. (“Grassmanality” is defined to be even or odd: a Grassmann even (odd) field is a commuting (anticommuting) field). The master action must obey the boundary condition of reducing to the classical action when the antifields are set to zero. (Note that in general the set of fields ϕ^s will be larger than the set of fields ϕ^i that appear in the classical action). Independence of the S -matrix on the gauge-fixing procedure is equivalent to the BV master equation

$$\frac{1}{2}\{S, S\} = -\hbar\Delta S \quad [26]$$

The antibracket $\{, \}$ and the Δ operator are defined as

$$\begin{aligned} \{A, B\} &\equiv \frac{\partial_r A}{\partial \phi^s} \frac{\partial_l B}{\partial \phi_s^*} - \frac{\partial_r A}{\partial \phi_s^*} \frac{\partial_l B}{\partial \phi^s} \\ \Delta &\equiv \frac{\partial_r}{\partial \phi^s} \frac{\partial_l}{\partial \phi_s^*} \end{aligned} \quad [27]$$

where ∂_l and ∂_r are derivatives from the left and from the right. It is often convenient to expand S in powers of \hbar , $S = S_0 + \hbar S_1 + \hbar^2 S_2 + \dots$, with

$$\begin{aligned} \{S_0, S_0\} &= 0 \\ \{S_0, S_1\} + \{S_1, S_0\} &= -2\hbar\Delta S_0, \dots \end{aligned} \quad [28]$$

With these definitions in place, we shall simply describe the answer, which is extremely elegant. In OSFT the full set of fields and antifields is packaged in a single string field $|\Phi\rangle$ of unrestricted ghost number. If we write

$$\begin{aligned} |\Phi\rangle &= |\Phi_- \rangle + |\Phi_+ \rangle \\ \text{with } G(\Phi_-) &\leq 1 \text{ and } G(\Phi_+) \geq 2 \end{aligned} \quad [29]$$

all the fields are contained in $|\Phi_- \rangle$ and all the antifields in $|\Phi_+ \rangle$. To make the pairing explicit, we pick a basis $\{|\Phi_s\rangle\}$ of $\mathcal{H}_{\text{BCFT}_0}$, and define a conjugate basis $\{|\Phi_s^C\rangle\}$ by

$$\langle \Phi_r^C | \Phi_s \rangle = \delta_{rs} \quad [30]$$

Clearly, $G(\Phi_s^C) + G(\Phi_s) = 3$. Then

$$|\Phi_- \rangle = \sum_{G(\Phi_s) \leq 1} |\Phi_s\rangle \phi^s, \quad |\Phi_+ \rangle = \sum_{G(\Phi_s) \geq 1} |\Phi_s^C\rangle \phi_s^* \quad [31]$$

Basis states $|\Phi_s\rangle$ with even (odd) ghost number $G(\Phi_s)$ are defined to be Grassmann even (odd). The full string field $|\Phi\rangle$ is declared to be Grassmann odd. It follows that ϕ^s is Grassmann even (odd) for $G(\Phi_s)$ odd (even), and that the corresponding antifield ϕ_s^* has the opposite Grassmanality of ϕ^s , as it must be. With this understanding of $|\Phi\rangle$, the classical master action S_0 is identical in form to the Witten action [16]! The boundary condition is satisfied; indeed, setting $|\Phi_+ \rangle = 0$, the ghost number anomaly implies that only the terms with $G = +1$ survive. The equation $\{S_0, S_0\} = 0$ follows from straightforward manipulations using the algebraic identities [17]. On the other hand, the issue of whether $\Delta S_0 = 0$, or whether instead quantum corrections are needed to satisfy full BV master equation, is more subtle and has never been fully resolved. The Δ operator receives singular contributions from the same region of moduli space responsible for the appearance of closed-string poles, which are discussed below. (See Thorn (1989) for a classic statement of this issue). It seems possible to choose a basis in $\mathcal{H}_{\text{BCFT}_0}$ such that there are no quantum corrections to S_0 (Erler and Gross 2004). In the following we shall derive the Feynman rules implied by S_0 alone.

SFT Diagrams and Minimal Area Metrics

Imposing the Siegel gauge condition $b_0\Phi=0$, one finds the gauge-fixed action

$$S_{\text{gf}} = -\frac{1}{g_o^2} \left(\frac{1}{2} \langle \Phi | c_0 L_0 | \Phi \rangle + \frac{1}{3} \langle \Phi | \Phi * \Phi \rangle + \langle \beta | b_0 | \Phi \rangle \right) \quad [32]$$

where β is a Lagrangian multiplier. The propagator reads

$$\frac{b_0}{L_0} = b_0 \int_0^\infty dT e^{-TL_0} \quad [33]$$

Since L_0 is the first-quantized open-string Hamiltonian, e^{-TL_0} is the operator that evolves the open-string wave functions $\Psi[X(\sigma)]$ by Euclidean world sheet time T . It can be visualized as a flat rectangular strip of “horizontal” width π and “vertical” height T . Each propagator comes with an antighost insertion

$$b_0 = \int_0^\pi b(\sigma) \quad [34]$$

integrated on a horizontal trajectory.

The only elementary interaction vertex is the mid-point three-string overlap, visualized in Figure 3. We are instructed to draw all possible diagrams with given external legs (represented as semi-infinite strips), and to integrate over all Schwinger parameters $T_i \in [0, \infty)$ associated with the internal propagators. The claim is that this prescription reproduce precisely the first-quantized result [1]. This follows if we can show that (1) the OSFT Feynman rules give a unique cover of the moduli space of open Riemann surfaces; (2) the integration measure agrees with the measure $[d\mu_\alpha]$ in [1]. The latter property holds because the antighost insertion [34] is precisely the one prescribed by the Polyakov formalism for integrating over the moduli T_i . To show point (1), we introduce the concept of minimal-area metrics, which has proved very fruitful. (Here and below, our discussion of

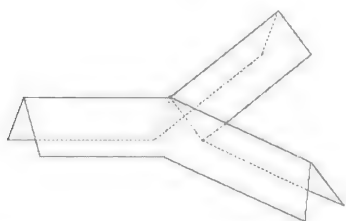


Figure 3 The cubic vertex represented as the mid-point gluing of three strips.

minimal-area metrics will summarize ideas developed mainly by Zwiebach.) Quite generally, the Feynman rules of an SFT provide us with a cell decomposition of the appropriate moduli space of Riemann surfaces, a way to construct surfaces in terms of vertices and propagators. Given a Riemann surface (for fixed values of its complex moduli), the SFT must associate with it one and only one string diagram. The diagram has more structure than the Riemann surface: it defines a metric on it. In all known covariant SFTs, this is the metric of minimal area obeying suitable length conditions. Consider the following:

Minimal-area problem for open SFT Let \mathcal{R}_o be a Riemann surface with at least one boundary component and possibly punctures on the boundary. Find the (conformal) metric of minimal area on \mathcal{R}_o such that all nontrivial Jordan open curves have length greater than or equal to π . (A curve is said to be nontrivial if it cannot be continuously shrunk to a point without crossing a puncture.)

An OSFT diagram (for fixed values of its T_i), defines a Riemann surface \mathcal{R}_o endowed with a metric solving this minimal-area problem. This is the metric implicit in its picture: flat everywhere except at the conical singularities of defect angle $(n-2)\pi$ when n propagators meet symmetrically. (For $n=3$, these are the elementary cubic vertices; for $n>3$, they are effective vertices, obtained when propagators joining cubic vertices collapse to zero length.) It is not difficult to see both that the length conditions are obeyed, and that the metric cannot be made smaller without violating a length condition. Conversely, any surface \mathcal{R}_o endowed with a minimal-area metric, corresponds to an OSFT diagram. The idea is that the minimal-area metric must have open geodesics (“horizontal trajectories”) of length π foliating the surface. The geodesics intersect on a set of measure zero – the “critical graph” where the propagators are glued. Bands of open geodesics of infinite height are the external legs of the diagram, while bands of finite height are the internal propagators.

The single cover of moduli space is then ensured by an existence and uniqueness theorem for metrics solving the minimal-area problem for OSFT. These metrics are seen to arise from Jenkins–Strebel quadratic differentials. Existence shows that the Feynman rules of OSFT generate each Riemann surface \mathcal{R}_o at least once. Uniqueness shows that there is no overcounting: since different diagrams correspond to different metrics (by inspection of their picture), no Riemann surface can be generated twice.

Closed Strings in OSFT

As is familiar, the open-string S -matrix contains poles due to the exchange of on-shell open and closed strings. The closed-string poles are present in nonplanar loop amplitudes. We have seen that OSFT reproduces the standard S -matrix. Factorization over the open-string poles is manifest, it corresponds to propagator lengths T_i going to infinity. Surprisingly, the closed-string poles are also correctly reproduced, despite the fact that OSFT treats only the open strings as fundamental dynamical variables. In some sense, closed strings must be considered as derived objects in OSFT. Factorizing the amplitudes over the closed-string poles, one finds that on-shell closed-string states can be represented, at least formally, as certain singular open-string fields with $G = +2$, closely related to the (formal) identity string field. The picture is that of a folded open string, whose left and right halves precisely overlap, with an extra closed-string vertex operator inserted at the mid-point. The corresponding open/closed vertex is given by

$$\langle \Psi_{\text{phys}} | \Phi \rangle_{\text{OC}} \equiv \langle \Psi_{\text{phys}}(0) \mathcal{I} \circ \Phi(0) \rangle_D$$

$$\mathcal{I} = \left(\frac{1 + iz}{1 - iz} \right)^2 \quad [35]$$

and describes the coupling to the open-string field of a nondynamical, on-shell closed string $|\Psi_{\text{phys}}\rangle$. It is possible to add this open/closed vertex to the OSFT action. Remarkably, the resulting Feynman rules give a single cover of the moduli space of Riemann surfaces with at least one boundary, with open and closed punctures. This is shown using the same minimal-area problem as above, but now allowing for surfaces with closed punctures as well.

We should finally mention that the structure of OSFT emerges frequently in topological string theory, in contexts where open/closed duality plays a central role. Two examples are the interpretation of Chern–Simons theory as the OSFT for the A -model on the conifold, and the interpretation of the Kontsevich matrix integral for topological gravity as the OSFT on FZZT branes in $(2,1)$ minimal string theory.

Closed Bosonic SFT

The generalization to covariant closed SFT is nontrivial, essentially because the requisite closed-string decomposition of moduli space is much more complicated.

The free theory parallels the open case, with a minor complication in the treatment of the CFT zero

modes. The closed-string field is taken to live in a subspace of the matter + ghost state space, $|\Psi\rangle \in \tilde{\mathcal{H}}_{\text{CFT}_0}$, where the tilde means that we impose the subsidiary conditions

$$b_0^- |\Psi\rangle = L_0^- |\Psi\rangle = 0, \quad b_0^- \equiv b_0 - \bar{b}_0, \quad [36]$$

$$L_0^- \equiv L_0 - \bar{L}_0$$

In the classical theory, the string field carries ghost number $G = +2$, since it is the off-shell extension of the familiar closed-string physical states, and the quadratic action reads

$$S \sim \langle \Psi, Q_c \Psi \rangle \quad [37]$$

Here Q_c is the usual closed BRST operator. The inner product \langle, \rangle is defined in terms of the BPZ inner product, with an extra insertion of $c_0^- \equiv c_0 - \bar{c}_0$,

$$\langle A, B \rangle \equiv \langle A | c_0^- | B \rangle \quad [38]$$

In [37] $G_{\text{top}} = +6$, as it should be. Without the extra ghost insertion and the subsidiary conditions [36] it would not be possible to write a quadratic action. The linearized equations of motion and gauge invariance,

$$Q_c |\Psi\rangle = 0, \quad |\Psi\rangle \sim |\Psi\rangle + Q_c |\Lambda\rangle, \quad |\Lambda\rangle \in \tilde{\mathcal{H}}_{\text{CFT}_0}^{(1)} \quad [39]$$

give the expected cohomological problem. The fact that the cohomology is computed in the semirelative complex, $b_0^- |\Psi\rangle = b_0^- |\Lambda\rangle = 0$, well known from the operator formalism of the first-quantized theory, is recovered naturally in the second-quantized treatment.

The interacting action is constructed iteratively, by demanding that the resulting Feynman rules give a (unique) cover of moduli space. This requires the introduction of infinitely many elementary string vertices $\mathcal{V}_{g,n}$, where n is the number of closed-string punctures and g the genus. This decomposition of moduli space is more intricate than the decomposition that arises in OSFT, but is in fact analogous to it, when characterized in terms of the following.

Minimal-area problem for closed SFT Let \mathcal{R}_c be a closed Riemann surface, possibly with punctures. Find the (conformal) metric of minimal area on \mathcal{R} such that all nontrivial Jordan closed curves have length greater than or equal to 2π .

The minimal-area metric induces a foliation of \mathcal{R}_c by closed geodesics of length 2π . In the classical theory ($g=0$), the minimal-area metrics arise from Jenkins–Strebel quadratic differentials (as in the open case), and geodesics intersect on a measure-zero set. For $g > 0$, however, there can be foliation bands of geodesics that cross. By staring at the foliation, we can break up the surface into vertices and propagators. In correspondence with each puncture, there is a band of

infinite height, a flat semi-infinite cylinder of circumference 2π , which we identify as an external leg of the diagram. We mark a closed geodesic on each semi-infinite cylinder, at a distance π from its boundary. Bands of finite height (internal bands not associated to punctures) correspond to propagators if their height is greater than 2π , otherwise they are considered part of an elementary vertex. Along any internal cylinder of height greater than 2π , we mark two closed geodesics, at a distance π from the boundary of the cylinder. If we now cut open all the marked curves, the surface decomposes into a number of semi-infinite cylinders (external legs), finite cylinders (internal propagators) and surfaces with boundaries (elementary interactions). Each elementary interaction of genus g and with n boundaries is an element of $\mathcal{V}_{g,n}$. A crucial point of this construction is that we took care of leaving a “stub” of length π attached to each boundary. Stubs ensure that sewing of surfaces preserves the length condition on the metric (no closed curve shorter than 2π).

These geometric data can be translated into an iterative algebraic construction of the full quantum action $S[\Psi]$. The $\mathcal{V}_{g,n}$ satisfy geometric recursion relations whose algebraic counterpart is the quantum BV master equation for $S[\Psi]$. Remarkably, the singularities of the Δ operator encountered in OSFT are absent here, precisely because of the presence of the stubs. We refer to Zwiebach (1993) for a complete discussion of closed SFT.

Open/Closed SFT

There is also a covariant SFT that includes both open and closed strings as fundamental variables. The Feynman rules arise from the following problem.

Minimal-area problem for open/closed SFT Let \mathcal{R}_{oc} be a Riemann surface, with or without boundaries, possibly with open and closed punctures. Find the (conformal) metric of minimal area on \mathcal{R}_{oc} such that all nontrivial Jordan open curves have length greater than or equal to $l_o = \pi$, and all nontrivial Jordan closed curves have length greater than or equal to $l_c = 2\pi$.

The surface \mathcal{R}_{oc} is decomposed in terms of elementary vertices $\mathcal{V}_{b,m}^{g,n}$ (of genus g , b boundary components, n closed-string punctures and m open-string punctures) joined by open and closed propagators. Degenerations of the surface correspond always to propagators becoming of infinite length – factorization is manifest both in the open and in the closed channel.

The SFT described in the section “Closed strings in OSFT” (Witten OSFT augmented with the single

open/closed vertex [35]) corresponds to taking $l_o = \pi$ and $l_c = 0$. Varying $l_c \in [0, 2\pi]$, we find a whole family of interpolating SFTs. This construction clarifies the special status of the Witten theory: moduli space is covered by a single cubic open overlap vertex, with no need to introduce dynamical closed strings, but at the price of a somewhat singular formulation.

Classical Solutions in Open SFT

In the present formulation of SFT, a background (a classical solution of string theory) must be chosen from the outset. The very definition of the string field requires to specify a (B)CFT₀. Intuitively, the string field lives in the “tangent” to the “theory space” at a specific point – where “theory space” is some notion of a “space of 2D (boundary) quantum field theories,” not necessarily conformal. In the early 1990s independence from the choice of background was demonstrated for infinitesimal deformations: the SFT actions written using neighboring (B)CFTs are indeed related by a field redefinition. In recent years, it has become apparent that at least the open-string field reaches out to open-string backgrounds a finite distance away – possibly covering the whole of theory space. (Classical solutions of closed SFT are beginning to be investigated at the time of this writing (2005)).

The OSFT action written using BCFT₀ data is just the full world volume action of the D-brane with BCFT₀ boundary conditions. Which classical solutions should we expect in this OSFT? In the bosonic string, Dp branes carry no conserved charge and are unstable. This instability is reflected in the presence of a mode with $m^2 = -1/\alpha'$, the open-string tachyon $T(x^\mu)$, $\mu = 0, \dots, p$. From this physical picture, Sen argued that:

1. the tachyon potential, obtained by eliminating the higher modes of the string field by their equations of motion, must admit a local minimum corresponding to the vacuum with no D-brane at all (henceforth, the tachyon vacuum, $T(x^\mu) = T_0$);
2. the value of the potential at T_0 (measured with respect to the BCFT₀ point $T=0$) must be exactly equal to minus the tension of the brane with BCFT₀ boundary conditions;
3. there must be no perturbative open-string excitations around the tachyon vacuum; and
4. there must be space-dependent “lump” solutions corresponding to lower-dimensional branes. For example, a lump localized along one world volume direction, say x^1 , such that $T(x^1) \rightarrow T_0$ as $x^1 \rightarrow \pm\infty$, is identified with a D($p-1$) brane.

Sen's conjectures have all been verified in OSFT. (See Sen (2004) and Taylor and Zwiebach (2003) for reviews). The deceptively simple-looking equations of motion (in Siegel gauge)

$$L_0|\Phi\rangle + b_0(|\Phi\rangle * |\Phi\rangle) = 0 \quad [40]$$

are really an infinite system of coupled equations, and no analytic solutions are known. Turning on a vacuum expectation value (VEV) for the tachyon drives into condensation an infinite tower of modes. Fortunately, the approximation technique of “level truncation” is surprisingly effective. The string field is restricted to modes with an L_0 eigenvalue smaller than a prescribed maximal level L . For any finite L , the truncated OSFT contains a finite number of fields and numerical computations are possible. Numerical results for various classical solutions converge quite rapidly as the level L is increased.

The most important solution is the string field $|T\rangle$ that corresponds to the tachyon vacuum. A remarkable feature of $|T\rangle$ is universality: it can be written as a linear combination of modes obtained by acting on the tachyon $c_1|0\rangle$ with ghost oscillators and matter Virasoro operators,

$$|T\rangle = T_0 c_1|0\rangle + u L_{-2}^m c_1|0\rangle + v c_{-1}|0\rangle + \dots$$

This implies that the properties of $|T\rangle$ are independent of any detail of BCFT₀, since all computations involving $|T\rangle$ can be reduced to purely combinatorial manipulations involving the ghosts and the Virasoro algebra. The numerical results strongly confirm Sen's conjectures, and indicate that the tachyon vacuum is located at a non-singular point in configuration space. Numerical solutions describing lower-dimensional branes and exactly marginal deformations are also available. For example, the full family of solutions interpolating between a D1 and a D0 brane at the self-dual radius has been found. There is increasing evidence that the open-string field provides a faithful map of the open-string landscape.

Vacuum SFT: D-branes as Projectors

In the absence of a closed-form expression for $|T\rangle$, we are led to guesswork. When expanded around $|T\rangle$, the OSFT is still cubic, only with a different kinetic term \mathcal{Q} ,

$$S = -\kappa_0 \left[\frac{1}{2} \langle \Phi | \mathcal{Q} | \Phi \rangle + \frac{1}{3} \langle \Phi | \Phi * \Phi \rangle \right] \quad [41]$$

The operator \mathcal{Q} must obey all the formal properties [17], must be universal (constructed from ghosts and matter Virasoro operators), and must have trivial cohomology at $G = +1$. Another constraint comes

from requiring that [41] admits classical solutions in Siegel gauge. The choice

$$\begin{aligned} \mathcal{Q} &= \frac{1}{2i} (c(i) - \bar{c}(i)) \\ &= c_0 - (c_2 + c_{-2}) + (c_4 + c_{-4}) - \dots \end{aligned} \quad [42]$$

satisfies all these requirements. The conjecture (Rastelli *et al.* 2001) is that, by a field redefinition, the kinetic term around the tachyon vacuum can be cast into this form. This “purely ghost” \mathcal{Q} is somewhat singular (it acts at the delicate string mid-point), and presumably should be regarded as the leading term of a more complicated operator that includes matter pieces as well. The normalization constant κ_0 is formally infinite. Nevertheless, a regulator (e.g., level truncation) can be introduced, and physical observables are finite and independent of the regulator. The vacuum SFT ([41]–[42]) appears to capture the correct physics, at least at the classical level. Taking a matter/ghost factorized ansatz

$$|\Phi_g\rangle \otimes |\Phi_m\rangle \quad [43]$$

and assuming that the ghost part is universal for all D-branes solutions, the equations of motion reduce to following equations for the matter part:

$$|\Phi_m\rangle * |\Phi_m\rangle = |\Phi_m\rangle \quad [44]$$

A solution $|\Phi_m\rangle$ can be regarded as a projector acting in “half-string space.” Recall that the $*$ -product looks formally like a matrix multiplication [19]: the matrices are the string fields, whose “indices” run over the half-string curves. These projector equations have been exactly solved by many different techniques (see Rastelli (2004) for a review). In particular, there is a general BCFT construction that shows that one can obtain solutions corresponding to any D-brane configuration, including multiple branes – the rank of the projector is the number of branes. A rank-one projector corresponds to an open-string functional which is left/right split, $\Phi[X(\sigma)] = F_L(X_L)F_R(X_R)$. There is also clear analogy between these solutions and the soliton solutions of noncommutative field theory. The analogy can be made sharper using a formalism that rewrites the open-string $*$ -product as the tensor product of infinitely many Moyal products. (See Bars (2002) and references therein).

It is unclear whether or not multiple-brane solutions (should) exist in the original OSFT – they are yet to be found in level truncation. Understanding this and other issues, like the precise role of closed strings in the quantum theory seems to require a precise characterization of the allowed

space of open-string functionals. In principle, the path integral over such functionals would define the theory at the full nonperturbative level. This remains a challenge for the future.

Note Added in Proof Very recently, M Schnabl, building on previous work on star algebra projectors and related surface states (Rastelli L (2004) and references therein) was able to find the exact solution for the universal tachyon condensate in OSFT. This breakthrough is likely to lead to rapid new developments in SFT.

See also: Boundary Conformal Field Theory; BRST Quantization; Chern–Simons Models: Rigorous Results; Fedosov Quantization; The Jones Polynomial; Large- N and Topological Strings; Large- N Dualities; Noncommutative Geometry from Strings; Noncommutative Tori, Yang–Mills, and String Theory; Operads; Superstring Theories; Topological Quantum Field Theory: Overview; Two-Dimensional Conformal Field Theory and Vertex Operator Algebras.

Further Reading

- Bars I (2002) *MSFT: Moyal Star Formulation of String Field Theory*. Proceedings of 3rd International Sakharov Conference on Physics, Moscow, Russia, 24–29 Jun, arXiv:hep-th/0211238.
- Berkovits N (2001) Review of open superstring field theory, arXiv:hep-th/0105230.
- Berkovits N, Okawa Y, and Zwiebach B (2004) WZW-like action for heterotic string field theory. *Journal of High Energy Physics* 0411: 038 (arXiv:hep-th/0409018).
- Bordes J, Chan HM, Nellen L, and Tsou ST (1991) Half string oscillator approach to string field theory. *Nuclear Physics B* 351: 441.

- Erler TG and Gross DJ (2004) Locality, causality, and an initial value formulation for open string field theory, arXiv:hep-th/0406199.
- Ohmori K (2001) *A Review on Tachyon Condensation in Open String Field Theories*. Master's thesis, University of Tokyo (arXiv:hep-th/0102085).
- Okawa Y (2002) Open string states and D-brane tension from vacuum string field theory. *Journal of High Energy Physics* 0207: 003 (arXiv:hep-th/0204012).
- Rastelli L (2004) Open string fields and D-branes. *Fortschritte der Physics*. 52: 302.
- Rastelli L, Sen A, and Zwiebach B *Vacuum String Field Theory*, Proceedings of the Strings 2001 Conference, TIFR, Mumbai, India, arXiv:hep-th/0106010.
- Schnabl M (2005) Analytic solution for tachyon condensation in open string theory. arXiv:hep-th/0511286.
- Sen A (2004) Tachyon dynamics in open string theory, arXiv:hep-th/0410103.
- Shatashvili SL *On Field Theory of Open Strings, Tachyon Condensation and Closed Strings*, Proceedings of the Strings 2001 Conference, TIFR, Mumbai, India, arXiv:hep-th/0105076.
- Siegel W (1988) Introduction To String Field Theory. *Advanced Series in Mathematical Physics* 8: 1–244 (arXiv:hep-th/0107094).
- Taylor W and Zwiebach B (2001) D-branes, tachyons, and string field theory. In: Gubser SS and Lykken JD (eds.) *Boulder 2001, Strings, Branes and Extra Dimensions, TASI 2001 Lectures*, pp. 641–759. Singapore: World Scientific.
- Thorn CB (1989) String field theory. *Physics Report* 175: 1.
- Witten E (1986) Noncommutative geometry and string field theory. *Nuclear Physics B* 268: 253.
- Zwiebach B (1993) Closed string field theory: quantum action and the B-V master equation. *Nuclear Physics B* 390: 33 (arXiv:hep-th/9206084).
- Zwiebach B (1998) Oriented open–closed string theory revisited. *Annals of Physics* 267: 193 (arXiv:hep-th/9705241).

String Theory: Phenomenology

A M Uranga, Consejo Superior de Investigaciones Científicas, Madrid, Spain

© 2006 Elsevier Ltd. All rights reserved.

String Theory and Compactification

The string theory provides a setup in which gauge and gravitational interactions can be described in a unified framework consistently at the quantum level. As such, it provides a candidate theory in which to describe the standard model of particle physics (describing quarks and leptons and their strong and electroweak interactions) and gravity within the same quantum theory.

The string theory has a unique fundamental scale M_s , fixed by the string tension, often encoded in the parameter α' of dimension $(\text{length})^{-2}$. All other scales are derived from this one and are background dependent.

Most of the string theory phenomenological model building has centered on the critical superstrings, which are ten dimensional (10D) and involve spacetime (as well as world-sheet) supersymmetry. There are five such different 10D theories: type IIA, type IIB, type I, and the $E_8 \times E_8$ and $SO(32)$ heterotic theories. The heterotic theories include nonabelian gauge fields and charged fermions in ten dimensions; hence, they constitute a promising setup to embed the standard model. On the other hand, the possibility of including D-branes

(which carry nonabelian gauge symmetries and charged matter) in compactifications of type II theories (and orientifolds thereof, like the type I theory itself) makes the latter reasonable alternative setups to embed the standard model as a brane world. The different 10D theories (as well as the 11D M-theory) are related by diverse dualities, also upon compactification. This suggests that they are just different limits of a unique underlying theory. For 4D models, this implies that the different classes of constructions are ultimately related by dualities, and that often a given model may be realized using different string theory constructions as starting points.

In order to recover 4D physics at low energies, compactification of the theory is required. In geometrical terms, the theory is required to propagate on a spacetime with geometry $M_4 \times X_6$, where M_4 is a 4D Minkowski space, and X_6 is a compact manifold. This description is valid in the regime of a large compactification volume, $\alpha'/R^2 \ll 1$ (where R is the overall scale of the compact manifold), where α' string theory corrections are negligible. Other 4D string models may be constructed using abstract conformal field theories. They may often be regarded as extrapolations of geometric compactifications to the regime of sizes comparable with the string length, where string theory corrections are relevant and the classical geometric picture does not hold.

In the simplest situation of geometrical compactification, not including additional backgrounds beyond the metric, the requirement of 4D spacetime supersymmetry (useful for the stability of the model, as well as of phenomenological interest) implies that the space X_6 is endowed with an $SU(3)$ holonomy metric. Existence of such metrics is guaranteed for Calabi–Yau spaces, namely Kähler manifolds with vanishing first Chern class.

There are a very large number of 4D supersymmetric string models that can be constructed using different starting string theories and different compactification manifolds. They lead to different 4D spectra, often including nonabelian gauge symmetries and charged chiral fermions (but only rarely resembling the actual standard model). In addition, for each given model, there exist, in general, a large number of massless 4D scalars, known as moduli, whose vacuum expectation values are not fixed. They parametrize different choices of the compactification data in a given topological sector (e.g., Kähler and complex structure moduli of the internal Calabi–Yau space). All physical parameters of the 4D theory vary continuously with the vacuum expectation values of these scalars.

All such models are on equal footing from the point of view of the theory. Hence, 4D string models suffer from a large arbitrariness. Although the breaking of supersymmetry clearly changes the picture qualitatively (e.g., flat directions associated to moduli are lifted by radiative corrections), it is difficult to evaluate this impact.

In this situation, most of the research in string theory phenomenology has centered on the study of generic properties of certain classes of compactifications, with the potential to lead to realistic structures (such as $N=1$ or no supersymmetry, nonabelian gauge symmetries with replicated sets of charged chiral fermions). Within each class, explicit models (as close as possible to the standard model) have also been constructed. Generic predictions or expectations for phenomenology can be obtained within each setup, but quantitative results, even for explicit models, are always functions of undetermined moduli vacuum expectation values. Tractable mechanisms for moduli stabilization are under active research, although only preliminary results are available presently.

The better-studied classes of models are compactifications of heterotic theories on Calabi–Yau spaces, and compactifications of type II theories (or orientifolds thereof) with D-branes. Other possibilities include the heterotic M-theory, the M-theory on G_2 holonomy varieties, the F-theory on Calabi–Yau 4-folds, etc. As already mentioned, different classes (or even explicit models) are often related by string duality.

Heterotic String Phenomenology

A large class of phenomenologically interesting string vacua, which has been explored in depth, is provided by 4D compactifications of (any of the two) perturbative heterotic string theories. Compactification on large volume manifolds can be described in the supergravity approximation. As described by Candelas, Horowitz, Strominger, and Witten, the requirement of 4D $N=1$ supersymmetry requires the internal manifold to be of $SU(3)$ holonomy, a condition which is satisfied by Calabi–Yau manifolds. In the presence of a curvature, the Bianchi identity for the Kalb–Ramond 2-form B is modified, so that, in general, it reads

$$dH = \text{tr } R^2 - \frac{1}{30} \text{tr } F^2 \quad [1]$$

where H is the field strength 3-form, R is the Ricci 2-form, and F is the field strength, in the adjoint representation, of the 10D gauge fields. Regarding the above equation in cohomology leads to a

consistency condition, forcing the background gauge bundle V to be topologically nontrivial, with

$$c_2(V) = c_2(TX_6) \quad [2]$$

where c_2 denotes the second Chern class, and TX_6 is the compactification tangent space.

The condition of supersymmetry implies that the gauge fields must be solutions of the Donaldson–Uhlenbeck–Yau equations. Existence of such a solution is guaranteed for holomorphic and stable gauge bundles. The simplest solution to these conditions is the so-called standard embedding, where the gauge connection is locally identical to the spin connection, but more general solutions exist and have been characterized for particular classes of Calabi–Yau manifolds (e.g., when they are elliptically fibered). The gauge background bundle V , with structure group H , breaks the 10D gauge symmetry G to its commutant subgroup G_{4D} . The latter corresponds to the 4D gauge symmetry. Moreover, the background bundle modifies the Kaluza–Klein reduction of the 10D charged fermions, leading to a nonzero number of replicated 4D chiral fermions. Decomposing the adjoint representation of G (in which 10D fermions transform) with respect to $G_{4D} \times H$,

$$\text{Adj } G = \bigoplus_i (R_{G_{4D},i}, R_{H,i}) \quad [3]$$

the net number of 4D chiral fermions in the representation $R_{G_{4D}}$ is given by the index of the Dirac operator coupled to V in the representation $R_{H,i}$. Condition [1] implies proper cancellation of chiral anomalies in the resulting theory. A simple and well-studied class is provided by standard embedding compactifications of the $E_8 \times E_8$ heterotic string theory, whose unbroken 4D gauge group is $E_6 \times E_8$. The number of families (i.e., chiral multiplets in the representation 27 of E_6) and conjugate families (in the $\overline{27}$) are given by the Hodge numbers

$$n_{27} = h_{1,1}(X_6), \quad n_{\overline{27}} = h_{2,1}(X_6) \quad [4]$$

More specifically, the harmonic representatives in each cohomology class represent the internal profile of the corresponding 4D fields. The net number of families is thus determined by the Euler characteristic $\chi(X_6)$

$$n_{\text{fam}} = |h_{1,1} - h_{2,1}| = \frac{1}{2} |\chi(X_6)| \quad [5]$$

Recently, much progress in heterotic model building has been achieved in nonstandard embedding compactifications by the detailed construction of holomorphic stable bundles and the computation of the diverse indexes. In particular, explicit models with just the minimal supersymmetric standard model spectrum have been constructed.

The above geometric approach has several limitations. On the technical side, the construction of explicit holomorphic and stable gauge bundles is nontrivial from the mathematical viewpoint. On the more fundamental side, it allows one to explore only the large volume limit of heterotic compactifications.

Further insight into the latter aspect can be obtained via constructions based on exactly solvable conformal field theories (CFTs), which describe the world-sheet string dynamics in compactifications, including all α' corrections, and, therefore, allowing one to enter the small volume regime. The simplest such compactifications are provided by toroidal orbifolds, which describe string propagation in quotients of toroidal compactifications by a discrete group Γ . From the world-sheet viewpoint, they are described by 2D free CFT, but which include sectors of closed strings with boundary conditions twisted by elements of Γ . The resulting 4D theory contains chiral fermions, arising from the untwisted and twisted sectors. In the former, the nonchiral spectrum of toroidal compactification suffers a projection onto the Γ -invariant states and leads to chirality. Twisted sectors are localized at the fixed points of the orbifold action, where the local supersymmetry is reduced, leading naturally to chiral fermions.

Many of these models can be regarded as limits of compactifications on Calabi–Yau spaces in the limit in which they become locally flat and develop conical singularities (and similarly, their gauge bundles become locally flat and with curvature localized near the singular points). Indeed, flat directions involving moduli fields in the twisted sector often exist, which correspond to geometric blow-ups of the singular point that resolve the conical singularities to yield a smooth Calabi–Yau.

The theories remain simple and solvable for any value of the untwisted moduli (namely moduli of the underlying toroidal compactification). This allows the discussion of their low-energy effective action including the explicit dependence on the untwisted moduli, while only partial results for the dependence on twisted moduli are known.

Other approaches, such as free fermion constructions or Gepner models, also provide exact descriptions of compactifications, although only at a point of the moduli space, deep inside the small volume regime.

Exact CFT constructions provide a small volume description of Calabi–Yau compactifications, at least for particular models. Moreover, their consistency conditions (modular invariance of the partition function) provide a stringy version of the large volume geometric condition implied by eqn [2]. The

constructions also show the existence of full-fledged string theory constructions with properties similar to geometric compactifications, but incorporating all α' corrections.

Within the general class of perturbative heterotic string models, a certain number of phenomenologically interesting statements are quite generic.

- The 4D Planck scale M_P and gauge couplings g_{YM} (at the string scale) are related to the fundamental string scale by

$$M_s = M_P g_{YM} \quad [6]$$

This implies that the string scale is close to the 4D Planck scale. In this situation, supersymmetry can stabilize the electroweak scale against radiative corrections.

- 4D heterotic models contain certain $U(1)$ symmetries, whose gauge bosons actually get Stueckelberg masses due to $B \wedge F$ couplings to components of the 2-form. Such $U(1)$'s would correspond to global symmetries, but are violated at tree level by α' nonperturbative effects, namely world-sheet instantons. Hence, no continuous global symmetries exist, even perturbatively, in these models. Proton decay might, however, be avoided by discrete global symmetries. In any event, even without such symmetries, the large fundamental scale suppresses the processes mediating proton decay. Thus, the proton lifetime is naturally larger than present experimental bounds.
- Gauge coupling constants for the different gauge factors in the standard model unify at the string scale. This agrees with extrapolation from their electroweak values, assuming the minimal supersymmetric standard model content between the electroweak and string scale, up to a mismatch of scales (by a factor of 20). The latter may be addressed in diverse ways, such as threshold corrections, intermediate scales, or in the heterotic M-theory.
- Yukawa couplings are, in principle, computable. Explicit computations have been carried out in standard embedding geometric compactifications (where they amount to the overlap integral of the internal profiles of the 4D fields, namely a topological intersection number), and in orbifold models. They are in general moduli dependent, so their quantitative analysis is involved. Qualitatively, however, interesting patterns, such as hierarchical structures, are possible, for example, in specific orbifold models.

Heterotic models have been studied beyond the perturbative regime. For instance, the construction

of compactifications including nonperturbative objects, namely 5-branes, has been pursued; so has been the strong coupling limit of the $E_8 \times E_8$ heterotic, described by compactifications of the M-theory on an interval (the so-called heterotic M-theory or Horava–Witten theory). The strong coupling phenomena of the $SO(32)$ heterotic theory can be addressed using dual type I (or other type II orientifold) constructions.

D-Brane Phenomenology

A different setup for realistic string theory compactifications, within the so-called brane-world constructions, is provided by compactifications of type II string theories containing D-branes, or quotients thereof. A particularly relevant class of quotients involves quotienting out by world-sheet parity, accompanied by some Z_2 geometric action. The resulting theories are denoted type II orientifolds, and contain orientifold planes, subspaces fixed under the geometric action, corresponding to regions where the orientation of a string can flip. Type II compactifications with D-branes filling the noncompact dimensions must satisfy a set of consistency conditions, known as RR tadpole cancellation. This is the condition that, in the compact space, the charge of D-branes and orientifold planes under the different RR forms must cancel. For the Z -valued charges, the conditions read

$$\sum_a N_a Q_a + Q_{Op} = 0 \quad [7]$$

where N_a denotes the multiplicity of D-branes with charge vector Q_a under the RR fields, Q_{Op} is the charge vector of the orientifold planes. Additional discrete conditions may be present if the relevant K-theory group (classifying D-brane charges in the corresponding background) contains torsion pieces.

The most familiar example of these constructions is provided by the type I string theory, which is an orientifold quotient of the type IIB theory by world-sheet parity (with no geometric action). The model can be regarded as containing one orientifold 9-plane and 32 D9 branes (all filling out 10D spacetime), such that their RR charges with respect to the (nondynamical) RR 10-form cancel.

Supersymmetric geometric compactifications of type II theories and orientifolds must correspond to compactification on Calabi–Yau spaces in order to have a preserved spinor. Models with D-branes filling the noncompact dimensions may be broadly classified into two classes: type IIB compactifications with $D(3+2p)$ -branes, wrapped on holomorphic $2p$ -cycles, and carrying holomorphic and stable

world-volume gauge bundles, and type IIA compactifications with D6 branes wrapped on special Lagrangian 3-cycles (in general, models with D4 and D8 branes are not allowed since Calabi–Yau spaces do not have nontrivial 1- or 5-cycles on which to wrap the branes). This classification is a large volume realization of the general classification of supersymmetric configurations of D-branes into two classes, denoted A and B.

Intersecting Brane Worlds

Type IIA compactifications with A-branes correspond to compactifications of type IIA theory (or orientifolds thereof) with D6 branes wrapped on 3-cycles of the internal Calabi–Yau space. In these models, each stack of N D6 branes generically leads to a $U(N)$ gauge factor. Chirality arises from open strings stretched between pairs of branes at the corresponding intersections. The chiral fermions from an open string stretched between branes a and b transform in the bifundamental representation $(\square_a, \overline{\square}_b)$ of the gauge factors $U(N_a) \times U(N_b)$ of the intersecting D6 brane stacks. In general, two 3-cycles in a 6D manifold intersect at points of the internal space. Hence, such fermions arise in several families, whose (net) number is given by the (net) number of intersections of the corresponding 3-cycles Π_a, Π_b , namely the topological invariant intersection number of their homology classes

$$I_{ab} = [\Pi_a] \cdot [\Pi_b] \quad [8]$$

Simple modifications of the above rules arise in some sectors in the presence of orientifold planes (e.g., the reduction of the gauge symmetry from unitary to orthogonal or symplectic factors for branes on top of orientifold planes).

The RR tadpole cancellation conditions specify that the total homological charge carried by the D6 branes (and the orientifold 6-planes) cancel. They imply automatic cancellation of cubic nonabelian anomalies, and the cancellation of mixed $U(1)$ anomalies by a Green–Schwarz mechanism mediated by 4D scalars from the RR closed-string sector.

Explicit models with SM spectrum have been constructed in orientifolds of toroidal compactifications in the nonsupersymmetric case, and in orbifolds thereof in supersymmetric cases. The generalization of the above construction beyond toroidal situations is, in principle, possible, but difficult, due to the mathematically challenging task of constructing special Lagrangian submanifolds for general Calabi–Yau manifolds.

Certain phenomenologically interesting quantities, such as gauge couplings and their threshold

corrections, Yukawa couplings, and other diverse correlation functions have been computed in toroidal cases, where the corresponding correlators are computable exactly in α' . Particularly interesting is the computation of Yukawa couplings, or, in general, of couplings involving only fields at intersections. These couplings arise from open-string world-sheet instantons, namely disks with boundaries on the D-branes corresponding to those intersections.

Type IIB Orientifolds

Type IIB compactifications with B-type branes contain several familiar classes of 4D models, for instance, compactifications of type I string theory on smooth Calabi–Yau spaces (whose description may be carried out using the effective supergravity action, in close analogy with the heterotic compactifications). Compactifications of type I string theory on orbifolds can be regarded as a particular realization of this, easily described using exact CFTs (although from the viewpoint of the general description as B-branes, the appearance of lower-dimensional branes requires their mathematical description to involve coherent sheaves). Since open strings at orbifolds do not have twisted boundary conditions, chirality arises from the orbifold projection of the toroidally compactified theory on the spectrum.

Another example within this kind is provided by the so-called magnetized D-brane models. These correspond to toroidal compactifications of type I theory, with D9 branes carrying constant magnetic backgrounds for the internal components of the world-volume gauge fields. In this kind of model, although the closed-string sector is highly supersymmetric, the open-string spectrum has reduced supersymmetry, or no supersymmetry (if the bundle stability condition is relaxed). Chirality arises from the nontrivial index of the Dirac operator for open strings ending on D-branes with different world-volume magnetic fields. Explicit models have mainly centered on nonsupersymmetric models from orientifolds of T^6 , and on supersymmetric models from orientifolds of the $T^6/(Z_2 \times Z_2)$ orbifold. In both contexts, models with semirealistic spectra have been obtained: concretely nonsupersymmetric models with just the standard model spectrum, or supersymmetric models with the minimal supersymmetric standard model spectrum, plus nonchiral matter. Further, properties of the gauge coupling constants and the computation of the Yukawa couplings have been studied as functions of undetermined moduli.

Finally, a second large class of models constructed using B-type branes are given by lower-dimensional D-branes, for example, D3 branes, located at singular points in the internal compactification space. Since the massless sector of open strings is determined only in terms of the local structure of the singularity, these models have been mostly studied in noncompact setups. Resulting spectra can be encoded in quiver diagrams, related to those in the mathematical literature on the McKay correspondence. Semirealistic three-family models have been constructed based on systems of D3 and D7 branes at the C^3/Z_3 orbifold singularity.

Type IIB orientifold compactifications are also intimately related to F-theory compactifications on Calabi–Yau 4-folds, which provide a nonperturbative completion for such models.

Mirror symmetry exchanges type IIB and IIA compactifications with B- and A-type branes. Hence, it provides a map between the above two kinds of compactifications. This shows that type IIB orientifold models lead to spectra with structure similar to that of intersecting-branes worlds, and that they share many of their general properties.

As a particular example, toroidal models of intersecting D6 branes are mapped under mirror symmetry to models of magnetized D9 branes. This mirror map has been exploited to construct the same theories from both starting points and to recover certain quantities, such as the α' -exact Yukawa couplings in the IIA picture from a purely classical (no α' corrections) computation in the mirror IIB model. This is a particular application of the general proposal of homological mirror symmetry in compactifications with branes.

Type II orientifold compactifications with D-branes have also been explored beyond the geometric regime, using exact CFTs to describe the (analog of the) internal space, and crosscap and boundary states to describe (the analogs of) orientifold planes and D-branes. Formal developments in the construction of the latter in Gepner models have been successfully applied to obtain large classes of semirealistic 4D string models in this setup.

As compared with heterotic compactifications, the setup of D-brane models leads to several generic features:

- Since gauge sectors are localized on D-branes, and have a dilaton dependence different from gravitational interactions, the relation between the fundamental string scale and the 4D Planck scale and gauge coupling reads

$$M_{\text{P}}^2 g_{\text{YM}}^2 = \frac{M_{\text{s}}^{11-p} V_{\text{T}}}{g_{\text{s}}} \quad [9]$$

where V_{T} is a measure of the volume in the directions transverse to the brane, and g_{s} is the 10D string coupling. The above relation shows that it is possible to achieve large 4D Planck mass with a lower fundamental string scale by adjusting the transverse volume and the string coupling. This has been proposed by Antoniadis, Arkani-Hamed, Dimopoulos, and Dvali as an alternative to explain the Planck/weak hierarchy without supersymmetry.

- The compactifications contain several $U(1)$ gauge symmetries. For some of the corresponding gauge bosons, the 4D effective theory contains Stueckelberg masses of order M_{s} , due to $B \wedge F$ couplings to fields in the RR sector. These couplings make the $U(1)$ gauge bosons massive; hence, they are absent from the low-energy physics. Nevertheless, the $U(1)$'s remain as global symmetries exact in α' and to all orders in the perturbation theory in g_{s} . They are violated by D-brane instantons, which are nonperturbative in g_{s} . In many realistic models, the baryon number is one such global symmetry, and it prevents proton decay, even if the string scale is not large.
- In general, each gauge factor in the standard model arises from a different brane stack, and their gauge couplings at the string scale are controlled by different moduli. This implies that, generically, it is not natural to have gauge coupling unification in D-brane models. Particular models may enjoy enhanced discrete global symmetries at special points in moduli space where unification is achieved, thus making unification appear more natural in such examples. Similar statements apply for constructions which realize complete or partial unification of gauge groups at large scales (like string models of grand unification or of Pati–Salam type).
- As already mentioned, important quantities such as Yukawa couplings are, in principle, computable, although quantitative expressions have been derived only in a few examples, mostly in toroidal compactifications or quotients thereof. The results are moduli dependent, making it difficult to derive model-independent patterns.

M-Theory Phenomenology

Most of the phenomenological models from the M-theory have been constructed using the Horava–Witten theory (compactification of M-theory on S^1/Z_2) as starting point. This theory provides a description of the strong coupling regime of the $E_8 \times E_8$ heterotic theory, and many of its basic features are similar to those in the perturbative regime. In particular, the techniques used in model

building involve the construction of stable and holomorphic vector bundles and the computation of the relevant indexes to obtain the 4D gauge group and charge matter content. An important difference is that gauge interactions propagate only over the 10D boundaries of spacetime, while gravity propagates over the 11 dimensions. This makes the setup share some features of brane-world constructions, and, in particular, it allows one to lower the fundamental scale of the theory (the 11D Planck scales) to reconcile it with the traditional unification scale.

A different setup for M-theory phenomenology involves the compactification of the 11D theory on a 7-manifold of G_2 holonomy X_7 , in order to lead to $N=1$ supersymmetry in four dimensions. Although a fundamental formulation of the M-theory is lacking, duality arguments and indirect evidence can be used to show that nonabelian gauge symmetries of the A–D–E classical groups arise if X_7 contains 3-cycles of codimension-4 singularities, locally of the form C^2/Γ , with Γ an A–D–E Kleinian subgroup of $SU(2)$. Similarly, it can be shown that chiral multiplets charged under these gauge symmetries arise if X_7 contains certain codimension-7 singularities. The local geometry of the latter has been explicitly described, and can be regarded as lying at the intersections of codimension-4 singularities.

The direct construction of such singular G_2 holonomy manifolds is very difficult, and there are no known topological conditions that guarantee existence of such a metric for a fixed topology. However, the existence of large classes of such models can be indirectly shown by using duality arguments. Namely, any type IIA models of intersecting D6 branes and O6 planes, preserving $N=1$ supersymmetry, lifts to an M-theory compactification on a singular G_2 holonomy manifold. In fact, the local structure of the codimension-4 and -7 singularities agrees in particular cases with the local structure of D6 branes on 3-cycles and D6 brane intersections.

Further Topics

Some additional topics related to the phenomenology of the string theory, but not covered by the above model building description are discussed in the following.

Effective Actions

The construction of effective actions for such classes of models has been carried out in general in supersymmetric compactifications, using the

parametrization of the general 4D $N=1$ supergravity action in terms of the Kähler potential for the moduli and matter fields, the gauge kinetic functions, and the superpotential. The moduli action is quite universal, at least for geometric compactifications and for untwisted moduli in orbifold compactifications. For instance, the Kähler potential for the 4D dilaton multiplet S and the modulus T controlling the size of the internal manifold, in the large volume and weak coupling regime, reads

$$K = -\log(S + S^*) - 3\log(T + T^*) \quad [10]$$

The corresponding expression including matter fields is more model dependent, but known within each particular class.

Moduli Stabilization and Supersymmetry Breaking

Both issues are often related. Although moduli stabilization preserving supersymmetry is possible, it often occurs that the potential stabilizing moduli has its origin in mechanisms related to supersymmetry breaking.

The description of purely string theoretical mechanisms to break supersymmetry is difficult, and most approaches rely on field-theoretical mechanisms in the effective action. One of the better-studied mechanisms, mostly in the heterotic string setup (but also in type II compactifications), is gaugino condensation in a strongly coupled hidden sector, interacting with the standard model sector via gravitational (or perhaps additional gauge) interactions. Although explicit models with such hidden sectors and strong dynamics exist, they often result in runaway potentials for moduli. Racetrack scenarios where several condensates balance each other are possible but contrived.

A second mechanism to break supersymmetry, mostly explored in type IIB/F-theory compactifications, is the introduction of field-strength fluxes for p -form fields. Interestingly, such fluxes lead to nontrivial potentials depending on moduli, and generically breaking supersymmetry. The existence of several remnant flat directions in the leading α', g_s approximation, leaves unanswered the question of possible runaway moduli potentials in those directions. However, evidence for nonperturbative contributions stabilizing the remaining moduli at finite distance has been proposed. Preliminary results in the analysis of flux stabilized vacua have been obtained in simple examples of (still unrealistic) Calabi–Yau compactifications with small number of moduli.

Most explored mechanisms propose supersymmetry breaking below the Kaluza–Klein compactification

scale, and, therefore, can be described in the 4D effective theory. They can be nicely parametrized in terms of vacuum expectation values for the dilaton and geometric moduli of the compactification. This description allows for a computation of the soft terms using the expansion of the $N=1$ supergravity formulas in components. Concrete patterns, such as the universality of squark masses, or the complex phases of diverse soft terms, can be explored using this approach.

Alternative mechanisms of breaking supersymmetry at higher scales, such as the introduction of antibranes or nonsupersymmetric compactifications, lead to generic difficulties with stability.

Related to the question of supersymmetry breaking is the question of the cosmological constant. Unfortunately, there is no manifest mechanism in the string theory that explains the smallness of the observed value of this scale. Given that many aspects of both quantum gravity in the string theory and realistic model building (with proper supersymmetry breaking and moduli stabilization) are still under progress, an open-minded point of view on this problem and the proposed solutions is kept.

Cosmology

Although somewhat different from the traditional focus of string phenomenology, recent progress in observational cosmology has triggered much interest in string theory realizations of inflationary models (or alternatives such as pre-big bang scenarios). Most inflationary models have centered on using moduli as the inflaton field, due to their flat potentials. A simple setup in type II compactifications, known as brane inflation models, uses the modulus controlling a brane position as the inflaton field, which has a flat enough potential with a moderate fine-tuning. Such setups may lead to interesting additional features, such as a moderate but potentially observable density of cosmic strings created in the reheating process.

On the other hand, many interesting questions in string cosmology await further understanding of time-dependent backgrounds in the string theory.

Retrospect

It is remarkable that the formal framework of the string theory admits tractable solutions with reasonable resemblance to the structure of the

standard model. In particular, generic features such as nonabelian gauge symmetry and chirality, coupled to gravity, are generic in 4D compactifications. This is already a success. In addition, much progress has been made in the general description of the relevant mathematical tools, and physical mechanisms and ingredients involved in these vacua, as well as in the explicit construction of models with the standard model spectrum (or supersymmetric extensions of it). Yet, many questions remain open and much more work is needed in order to make contact with the physics observed in nature.

See also: Brane Worlds; Compactification of Superstring Theory; Cosmology: Mathematical Aspects; Superstring Theories.

Further Reading

- Acharya B and Witten E (2001) Chiral fermions from manifolds of $G(2)$ holonomy, hep-th/0109152.
- Aldazabal G, Ibáñez LE, Quevedo F, and Uranga AM (2000) D-branes at singularities: a bottom up approach to the string embedding of the standard model. *Journal of High Energy Physics* 0008: 002.
- Angelantonj C and Sagnotti A (2002) Open strings. *Physics Reports* 371: 1–150.
- Angelantonj C and Sagnotti A (2003) Open strings – erratum. *Physics Reports* 376: 339–405.
- Antoniadis I, Arkani-Hamed N, Dimopoulos S, and Dvali GR (1998) New dimensions at a millimeter to a Fermi and superstrings at a TeV. *Physics Letters B* 436: 257–263.
- Bachas C (1995) A way to break supersymmetry, hep-th/9503030.
- Blumenhagen R, Cvetič M, Langacker P, and Shiu G (2005) Toward realistic intersecting D-brane models, hep-th/0502005.
- Candelas P, Horowitz GT, Strominger A, and Witten E (1985) Vacuum configurations for superstrings. *Nuclear Physics B* 258: 46–74.
- Donagi R, He Y-H, Ovrut BA, and Reinbacher R (2004) The spectra of heterotic standard model vacua, hep-th/0411156.
- Green MB, Schwarz JH, and Witten E (1987) *Superstring Theory*. Cambridge Monographs On Mathematical Physics, vols. 1 and 2. Cambridge: Cambridge University Press.
- Ibáñez LE (1987) The search for a standard model $SU(3) \times SU(2) \times U(1)$ superstring: an introduction to orbifold constructions. *Seoul Sympos.* 1986, 46.
- Polchinski J (1998) *String Theory*. vols. 1 and 2. Cambridge: Cambridge University Press.
- Uranga AM (2003) Chiral four-dimensional string compactifications with intersecting D-branes. *Classical and Quantum Gravity* 20: S373–S394.
- Witten E (1996) Strong coupling expansion of Calabi–Yau compactification. *Nuclear Physics B* 471: 135–158.

String Topology: Homotopy and Geometric Perspectives

R L Cohen, Stanford University, Stanford, CA, USA

© 2006 Elsevier Ltd. All rights reserved.

String topology is a new field of study involving the geometric and algebraic topology of spaces of loops and paths in manifolds. The subject was initiated in the important work of Chas and Sullivan (1999) who uncovered previously unknown algebraic structure in the homology and equivariant homology of loop spaces. While the structure is purely topological, it was motivated by formalisms in quantum field theory and string theory. Since that time this subject has attracted the attention of many mathematicians, but one of the main lines of research continues to be motivated by the attempt to understand the relation between this structure (and its generalizations) with topological and conformal field theories.

In order to describe some of the recent advances in this field, we begin with some notation. Throughout this article M^n will denote a closed, n -dimensional, oriented manifold. LM will denote the free loop space,

$$LM = \text{Map}(S^1, M)$$

For $D_1, D_2 \subset M$ closed submanifolds, $\mathcal{P}_M(D_1, D_2)$ will denote the space of paths in M that start at D_1 and end at D_2 ,

$$\mathcal{P}_M(D_1, D_2) = \{\gamma : [0, 1] \rightarrow M, \gamma(0) \in D_1, \gamma(1) \in D_2\}$$

The paths and loops we consider will always be assumed to be piecewise smooth. Such spaces of paths and loops are well known to be infinite-dimensional manifolds, and roughly speaking, string topology is the study of the intersection theory in these manifolds.

Recall that for closed, oriented manifolds, there is an intersection pairing,

$$H_r(M) \times H_s(M) \rightarrow H_{r+s-n}(M)$$

which is defined to be Poincaré dual to the cup product,

$$H^{n-r}(M) \times H^{n-s}(M) \xrightarrow{\cup} H^{2n-r-s}(M)$$

The geometric significance of this pairing is that if the homology classes are represented by submanifolds, P^r and Q^s with transverse intersection, then the image of the intersection pairing is represented by the geometric intersection, $P \cap Q$.

The remarkable result of Chas and Sullivan says that even without Poincaré duality, there is an intersection type product

$$\mu : H_p(LM) \times H_q(LM) \rightarrow H_{p+q-n}(LM)$$

that is compatible with both the intersection product on $H_*(M)$ via the map $ev : LM \rightarrow M(\gamma \rightarrow \gamma(0))$, and with the Pontrjagin product in $H_*(\Omega M)$.

The construction of this pairing involves consideration of the diagram,

$$LM \xleftarrow{\gamma} \text{Map}(8, M) \xrightarrow{e} LM \times LM \quad [1]$$

Here $\text{Map}(8, M)$ is the mapping space from the figure 8 to M , which can be viewed as the subspace of $LM \times LM$ consisting of those pairs of loops that agree at the basepoint. $\gamma : \text{Map}(8, M) \rightarrow LM$ is the map on mapping spaces induced by the pinch map $S^1 \rightarrow S^1 \vee S^1$.

Chas and Sullivan constructed this pairing by studying intersections of chains in loop spaces. A more homotopy-theoretic viewpoint was taken by Cohen and Jones (2002) who viewed $e : \text{Map}(8, M) \rightarrow LM \times LM$ as an embedding, and showed there is a tubular neighborhood homeomorphic to a normal given by the pullback bundle, $ev^*(TM)$, where $ev : LM \rightarrow M$ is the evaluation map mentioned above. They then constructed a Pontrjagin–Thom collapse map whose target is the Thom space of the normal bundle, $\tau_e : LM \times LM \rightarrow \text{Map}(8, M)^{ev^*(TM)}$. Computing τ_e in homology and applying the Thom isomorphism defines an “umkehr map,”

$$e_! : H_*(LM \times LM) \rightarrow H_{*-n}(\text{Map}(8, M))$$

The Chas–Sullivan loop product is defined to be the composition

$$\begin{aligned} \mu_* &= \gamma_* \circ e_! : H_*(LM \times LM) \rightarrow H_{*-n}(\text{Map}(8, M)) \\ &\rightarrow H_{*-n}(LM) \end{aligned}$$

Notice that the umkehr map $e_!$ can be defined for a generalized homology theory h_* whenever one has a Thom isomorphism of the tangent bundle, TM , which is to say a generalized homology theory h_* for which the representing spectrum is a ring spectrum, and which supports an orientation of M .

By twisting the Pontrjagin–Thom construction by the virtual bundle $-TM$, one obtains a map of spectra,

$$\tau_e : LM^{-TM} \wedge LM^{-TM} \rightarrow \text{Map}(8, M)^{ev^*(-TM)}$$

where LM^{-TM} is the Thom spectrum of the pullback of the virtual bundle $ev^*(-TM)$. Now we can compose, to obtain a multiplication,

$$LM^{-TM} \wedge LM^{-TM} \xrightarrow{\tau_e} \text{Map}(8, M)^{ev_0^*(-TM)} \xrightarrow{\gamma} LM^{-TM}$$

The following was proved by Cohen and Jones (2002).

Theorem 1 *Let M be a closed manifold, then LM^{-TM} is a ring spectrum. If M is orientable the ring structure on LM^{-TM} induces the Chas–Sullivan loop product on $H_*(LM)$ by applying homology and the Thom isomorphism.*

The ring structure on the spectrum LM^{-TM} was also observed by Dwyer and Miller using different methods.

Cohen and Godin (2004) generalized the loop product in the following way. Observe that the figure 8 is homotopy equivalent to the pair of pants surface P , which we think of as a genus 0 cobordism between two circles and one circle.

Furthermore, **Figure 1** is homotopic to the diagram of mapping spaces,

$$LM \xleftarrow{\rho_{\text{out}}} \text{Map}(P, M) \xrightarrow{\rho_{\text{in}}} (LM)^2$$

where ρ_{in} and ρ_{out} are restriction maps to the “incoming” and “outgoing” boundary components of the surface P . So the loop product can be viewed as a composition,

$$\begin{aligned} \mu &= \mu_P \\ &= (\rho_{\text{out}})_* \circ (\rho_{\text{in}})_! : (H_*(LM))^{\otimes 2} \rightarrow H_*(\text{Map}(P, M)) \\ &\rightarrow H_*(LM) \end{aligned}$$

where using the figure 8 to replace the surface P can be viewed as a technical device that allows one to define the umkehr map $(\rho_{\text{in}})_!$.

In general if one considers a surface of genus g , viewed as a cobordism from p incoming circles to q outgoing circles, $\Sigma_{g,p+q}$, one gets a similar diagram (**Figure 2**)

$$(LM)^q \xleftarrow{\rho_{\text{out}}} \text{Map}(\Sigma_{g,p+q}, M) \xrightarrow{\rho_{\text{in}}} (LM)^p$$

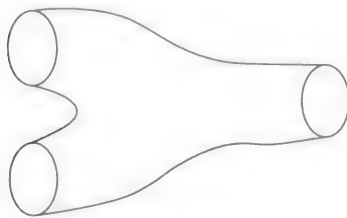


Figure 1 Pair of pants P .

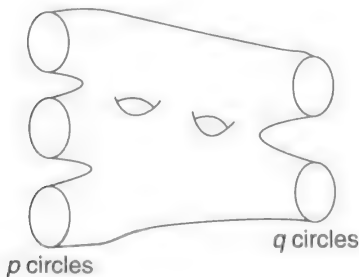


Figure 2 $\Sigma_{g,p+q}$.

Cohen and Godin (2004) used the theory of “fat” or “ribbon” graphs to represent surfaces as developed by Harer (1985), Penner (1987), and Strebel (1984), in order to define Pontrjagin–Thom maps,

$$\tau_{\Sigma_{g,p+q}} : (LM)^p \rightarrow \text{Map}(\Sigma_{g,p+q}, M)^{\nu(\Sigma_{g,p+q})}$$

where $\nu(\Sigma_{g,p+q})$ is the appropriately defined normal bundle of ρ_{in} . By applying (perhaps generalized) homology and the Thom isomorphism, they defined the umkehr map,

$$(\rho_{\text{in}})_! : H_*((LM)^p) \rightarrow H_{*+\chi(\Sigma_{g,p+q}) \cdot n}(\text{Map}(\Sigma_{g,p+q}, M))$$

where $\chi(\Sigma_{g,p+q}) = 2 - 2g - p - q$ is the Euler characteristic. Cohen and Godin then defined the string topology operation to be the composition,

$$\begin{aligned} \mu_{\Sigma_{g,p+q}} &= \rho_{\text{out}} \circ (\rho_{\text{in}})_! : H_*((LM)^p) \rightarrow H_{*+\chi(\Sigma_{g,p+q}) \cdot n} \\ &\times (\text{Map}(\Sigma_{g,p+q}, M)) \rightarrow H_{*+\chi(\Sigma_{g,p+q}) \cdot n}((LM)^q) \end{aligned}$$

They proved that these operations respect gluing of surfaces,

$$\mu_{\Sigma_1 \# \Sigma_2} = \mu_{\Sigma_2} \circ \mu_{\Sigma_1}$$

where $\Sigma_1 \# \Sigma_2$ is the glued surface as shown in **Figure 3**.

The coherence of these operations is summarized in the following theorem.

Theorem 2 (Cohen and Godin 2004). *Let h_* be any multiplicative generalized homology theory that supports an orientation of M . Then the assignment*

$$\Sigma_{g,p+q} \rightarrow \mu_{\Sigma_{g,p+q}} : h_*((LM)^p) \rightarrow h_*((LM)^q)$$

is a positive boundary topological quantum field theory. “Positive boundary” refers to the fact that the number of outgoing boundary components, q , must be positive.

A theory with open strings was initiated by Sullivan (2004) and developed further by A Ramirez (2005) and by Harrelson (2004). In this

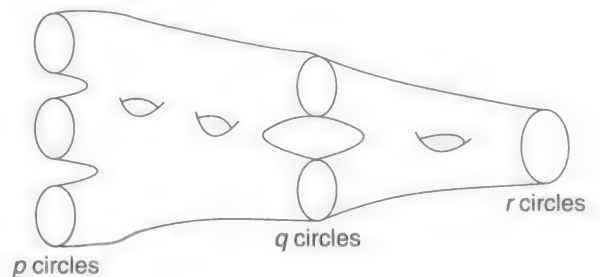


Figure 3 $\Sigma_1 \# \Sigma_2$.

setting one has a collection of submanifolds, $D_i \subset M$, referred to as “D-branes.” This theory studies intersections in the path spaces $P_M(D_i, D_j)$.

A theory with D-branes involves “open–closed cobordisms” which are cobordisms between compact one-dimensional manifolds whose boundary is partitioned into three parts:

- 1. Incoming circles and intervals.
- 2. Outgoing circles and intervals.
- 3. The rest is the “free boundary” which is itself a cobordism between the boundary of the incoming and boundary of the outgoing intervals. Each connected component of the “free boundary” is labeled by a D-brane (see Figure 4).

In a topological field theory with D-branes, one associates to each boundary circle a vector space V_{S^1} (in our case $V_{S^1} = H_*(LM)$) and to an interval whose endpoints are labeled by D_i, D_j , one associates a vector space V_{D_i, D_j} (in our case $V_{D_i, D_j} = H_*(P_M(D_i, D_j))$).

To an open–closed cobordism as above, one associates an operation from the tensor product of these vector spaces corresponding to the incoming boundaries to the tensor product of the vector spaces corresponding to the outgoing boundaries. Of course, these operations have to respect the relevant gluing of open–closed cobordisms.

By developing a theory of fat graphs that encode the open–closed boundary data, Ramirez was able to prove that there are string topology operations that form a positive boundary, topological quantum field theory with D-branes (Ramirez 2005).

We end these notes by a discussion of three applications of string topology to classifying spaces of groups.

Example 1 Application to Poincaré duality groups – (Abbaspour *et al.* to appear). For G any discrete

group, one has that the loop space of the classifying space satisfies

$$LBG \simeq \coprod_{[g]} BC_g$$

where $[g]$ is the conjugacy class determined by $g \in G$, and $C_g < G$ is the centralizer of g .

When BG is represented by a closed manifold, or more generally, when G is a Poincaré duality group, the Chas–Sullivan loop product then defines pairings among the homologies of the centralizer subgroups. Abbaspour *et al.* describe this loop product entirely in terms of group homology, thus giving structure to the homology of Poincaré-duality groups that previously had not been known.

Example 2 Applications to 3-manifolds. (Abbaspour 2005). Let $\iota: H_*M \rightarrow H_*(LM)$ be induced by inclusion of constant loops. This is a split injection of rings. Write $H_*(LM) = H_*(M) \oplus A_M$. We say $H_*(LM)$ has nontrivial extended loop products if the composition

$$A_M \otimes A_M \hookrightarrow H_*(LM) \otimes H_*(LM) \xrightarrow{\mu} H_*(LM)$$

is nontrivial.

Let M be a closed, irreducible 3-manifold. In a remarkable piece of work, Abbaspour showed the relationship between having a trivial extended loop product and M being “algebraically hyperbolic.” This means that M is a $K(\pi, 1)$ and its fundamental group has no rank-2 abelian subgroup. (If geometrization conjecture is true, this is equivalent to M admitting a complete hyperbolic metric.)

Example 3 The string topology of classifying spaces of compact Lie groups (Gruher (to appear) and of Gruher and Salvatore (to appear)). The goal of Gruher’s work is to construct string topological invariants of $LBG \simeq EG \times_G G$, where G acts on itself via conjugation. Ultimately, one would like to understand the relationship between this structure and the work of Freed (2003) on twisted equivariant K-theory, $K_G^{\tau}(G)$ and the Verlinde algebra.

The first observation in this program was to notice that the key ingredient in the forming of the Chas–Sullivan loop product is that the fibration $ev: LM \rightarrow M$ is a fiberwise monoid over a closed oriented manifold. The fiber is ΩM , which has the usual Pontrjagin product.

The following was proved by Gruher and Salvatore:

Lemma 3 *Let $G \rightarrow E \rightarrow M$ be a fiberwise monoid over a closed manifold M . Then E^{-TM} is a ring spectrum.*

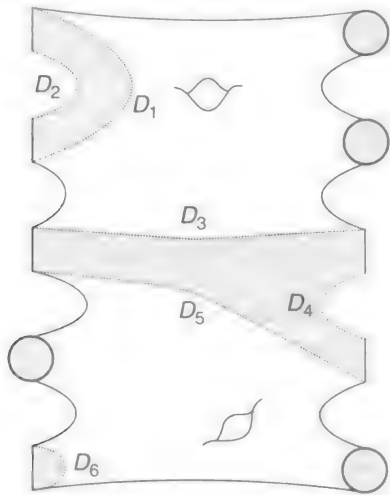


Figure 4 Open–closed cobordism.

The following construction gives a large supply of examples of such fiberwise monoids over manifolds.

Let $G \rightarrow P \rightarrow M$ be a principal G bundle over a closed manifold M . We can construct the corresponding adjoint bundle,

$$\mathrm{Ad}(P) = P \times_G G \rightarrow M$$

It is an easy observation that $G \rightarrow \mathrm{Ad}(P) \rightarrow M$ is a fiberwise monoid.

Theorem 4 $\mathrm{Ad}(P)^{-TM}$ is a ring spectrum. This ring structure is natural with respect to maps of principal G -bundles.

Let BG be classifying space of compact Lie groups. It is possible to construct a filtration of BG ,

$$M_1 \hookrightarrow M_2 \hookrightarrow \cdots \hookrightarrow M_i \subset M_{i+1} \hookrightarrow \cdots \hookrightarrow BG$$

where the M_i 's are compact, closed manifolds. An example of this is filtering $BU(n)$ by Grassmannians.

Let $G \rightarrow P_i \rightarrow M_i$ be the restriction of $EG \rightarrow BG$. By the above theorem one obtains an inverse system of ring spectra

$$P_1^{-TM_1} \leftarrow P_2^{-TM_2} \leftarrow \cdots \leftarrow P_i^{-TM_i} \leftarrow P_{i+1}^{-TM_{i+1}} \leftarrow \cdots$$

Theorem 5 The homotopy type of this pro-ring-spectrum is a well-defined invariant of BG . It is referred to as the “string topology of BG .”

Potential Application: Twisted K -theory and the Verlinde Algebra

Let G be a connected, compact Lie group. Using the observation that the loop space of a classifying space is the classifying space of the loop group, $L(BG) \simeq B(LG)$, the string topology gives new structure on the classifying space of these loop groups. In particular, one has new structure on the K -theory of these classifying spaces. Now classical results of Atiyah and Segal suggest that K -theory of classifying spaces should be related to the representation theory of the group. In this case, the representation theory of loop groups has been widely studied and is very important in conformal field theory.

Understanding the precise relationship between the string topology of the classifying space and this representation theory is an interesting area of current research. To motivate this, first recall that the loop space, LBG , has a well-known description as

$$LBG \simeq EG \times_{\mathrm{Ad}G}$$

where the right-hand side refers to the homotopy orbit space of the conjugation (or adjoint) action of G on itself. Thus, the homology $H_*(LBG)$ is the equivariant homology $H_*^G(G)$. Similarly, the

K -theory $K^*(LBG)$ maps to the equivariant K -theory, $K_G^*(G)$. Now in recent work of Freed (2003) twisted equivariant K -homology, $K_G^\tau(G)$ was shown to be isomorphic to the Verlinde algebra. This algebra is a space of representations of the loop group, LG . The multiplication in this algebra is the “fusion product,” coming from conformal field theory. One topic of current research is to understand the relationship between multiplicative structure coming from the string topology of BG , and this fusion product in the Verlinde algebra. More generally, the goal is to bring to bear the considerable calculational techniques of algebraic topology that are available in string topology, to understand the recently uncovered field theoretic structure of twisted K -theory (Freed 2003), and its applications to string theory.

Acknowledgment

The author was partially supported by a grant from the NSF.

See also: Mathematical Knot Theory; Topological Defects and Their Homotopy Classification.

Further Reading

- Abbaspour H (2005) On string topology of three manifolds. *Topology* 44: 1059–1091.
- Abbaspour H, Cohen RL, and Gruher K *String Topology of Poincaré Duality Groups* (in preparation).
- Chas M and Sullivan D (1999) String topology. To appear in *Annals of Mathematics*. Preprint: math.GT/9911159.
- Cohen RL and Godin V (2004) A polarized view of string topology. In: *Topology, Geometry, and Quantum Field Theory*, London Math. Soc. Lecture Notes, vol. 308, pp. 127–154.
- Cohen RL and Jones JDS (2002) A homotopy theoretic realization of string topology. *Mathematisches Annalen* 324: 773–798.
- Freed D, Hopkins M, and Teleman C Twisted K -theory and loop group representations. Preprint: math.AT/0312155.
- Godin V (2004) *A Category of Bordered Fat Graphs and the Mapping Class Group of a Bordered Surface*. Ph.D. thesis, Stanford University.
- Gruher K Ph.D. thesis, Stanford University (in preparation).
- Gruher K and Salvatore P *String Topology of Classifying Spaces* (in preparation).
- Harrelson E (2004) On the homology of open–closed string theory. Preprint: math.AT/0412249.
- Harer JL (1985) Stability of the homology of the mapping class groups of orientable surfaces. *Annals of Mathematics* 121: 215–249.
- Penner R (1987) The decorated Teichmüller space of punctured surfaces. *Communications in Mathematical Physics* 113: 299–339.
- Ramirez A (2005) *Open–Closed String Topology*. Ph.D. thesis, Stanford University.
- Strebel K (1984) *Quadratic Differentials*. Berlin: Springer.
- Sullivan D (2004) Open and closed string field theory interpreted in classical algebraic topology. In: Tillman U (ed.) *Topology, Geometry, and Quantum Field Theory*, London Math. Soc. Lecture Notes, vol. 308, pp. 344–357. Cambridge: Cambridge University Press.

Superfluids

D Einzel, Bayerische Akademie der Wissenschaften, Garching, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Superfluidity has been known to exist since the 1930s. This widespread phenomenon occurs in many-particle Bose and Fermi systems as different as liquid ^4He , liquid ^3He , atomic gases like Rb and Li, atomic nuclei, pulsars and last, but not least, in metals, where the itinerant electrons may become superfluid. This article is devoted to a unifying theoretical description of Bose and Fermi superfluidity. The mechanisms leading to superfluidity include Bose–Einstein condensation (BEC) and Bardeen, Cooper, and Schrieffer (BCS)–Leggett pairing correlations. We hope to be able to demonstrate why this fascinating phenomenon is – even roughly 80 years after its experimental discovery and its first theoretical explanation – still a subject of intensive research.

The phenomenon of superfluidity is closely connected with the apparent lack of any measurable flow resistance, which scales with the shear viscosity of the fluid. Its complete absence implies that the system is frictionless moving with zero viscosity. The observation of superfluidity is usually precluded by the solidification of most liquids as the temperature is lowered. Only systems with particularly light atoms (like the helium isotopes ^4He and ^3He) stay liquid down to the lowest temperatures. These systems are referred to as “quantum liquids,” since their liquid state is caused by the quantum-mechanical zero-point motion of the atoms. It should be noted that the Helium isotopes belong to two different kinds of elementary particles which can be distinguished by their statistics: ^4He is a spin-0 boson and ^3He a spin-1/2 fermion.

In 1924, Satyendra Nath Bose and Albert Einstein proposed that below a characteristic degeneracy temperature T_B , a macroscopic number of bosons can condense into the state of lowest energy $\epsilon_k = 0$. In the 1930s, Fritz London and Heinz London showed that this so-called Bose–Einstein condensate can be described by a macroscopic quantum-mechanical wave function like the one for a single elementary particle, but with the probability density replaced by the density of the condensed particles. By the end of the 1930s, the experimental results of Allen, Kamerlingh–Onnes, Keesom, Kapitza,

Miesener, Wolfke, and others accumulated the evidence that liquid ^4He undergoes a second-order phase transition at $T_\lambda = 2.17\text{ K}$ to a state referred to as a superfluid, since the liquid could flow without any sign of a flow resistance. This superfluid state was interpreted in terms of Bose condensation of the ^4He atoms in the liquid (London 1938).

In Figure 1 the P – T phase diagram of liquid ^4He is shown with a normal liquid phase, a solid phase and the superfluid phase below the λ -line at about 2 K.

Fermions cannot condense in a way similar to the BEC, due to the Pauli exclusion principle. In 1957 Bardeen, Cooper, and Schrieffer came up with their ingenious proposal that the superfluidity of the electron system (usually referred to as superconductivity) comes about through the formation of fermion pairs (quasibosons) in k -space in a spin-singlet state. In 1971, several superfluid phases of liquid ^3He at a few mK were discovered by Lee, Osheroff, and Richardson at Cornell University. Experimental aspects connected with the spin degrees of freedom of the quantum liquid gave strong evidence for Cooper pairing of the ^3He atoms in a spin-triplet state. In Figure 2 the zero-field P – T phase diagram of liquid ^3He is shown with a normal (Fermi) liquid phase, a solid phase and the superfluid A and B phases.

Immediately after this discovery, Anthony J Leggett applied the BCS ideas to liquid ^3He and introduced a generalized scheme, that allowed for triplet-pairing correlations. His theory turned out to describe a large variety of experimental results accurately. A new and exciting development set in when Bose–Einstein condensates were discovered for the first time in dilute gases of alkali atoms in 1995 by Cornell and Wiemann *et al.* (Rb), Ketterle *et al.* (Na), and Hulet *et al.* (Li).

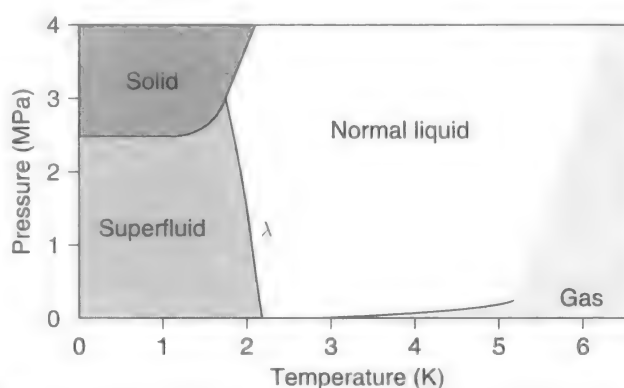


Figure 1 The phase diagram of liquid ^4He . Courtesy of Erkki Thuneberg.

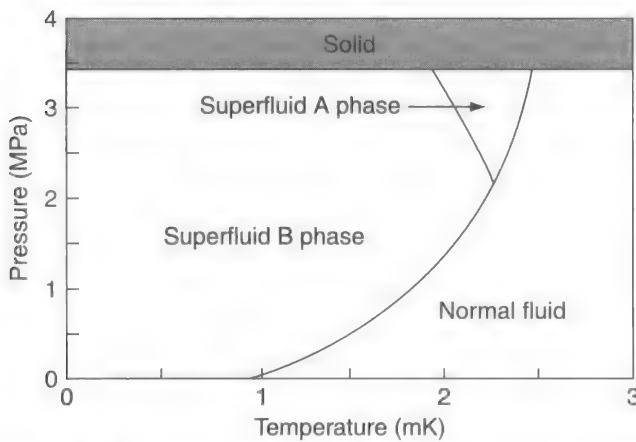


Figure 2 The phase diagram of liquid ^3He . Courtesy of Erkki Thuneberg.

Boson and Fermion Degeneracy

In what follows, the energy dispersion of Bose and Fermi systems is denoted as ϵ_k (free bosons/fermions would be represented by $\epsilon_k = \hbar^2 k^2 / 2m$). A large number of bosons can occupy Bose quantum states $|k\rangle$, the average occupation is dictated by the Bose-Einstein distribution

$$n_k = \frac{1}{e^{(\epsilon_k - \mu)/k_B T} - 1} \quad [1]$$

For Bose systems, the chemical potential is negative $\mu = -k_B T \alpha$ and α is fixed by the condition

$$n = \frac{1}{V} \sum'_k n_k = \frac{1}{\lambda_T^3} B_{\frac{3}{2}}(\alpha) \quad [2]$$

where the prime indicates the summation over excited states $|k| > 0$. In [2], $\lambda_T = h / \sqrt{2\pi m k_B T}$ denotes the thermal de Broglie wavelength which provides a criterion for the importance of quantum effects or degeneracy through $n \lambda_T^3 \geq O(1)$. The Bose integrals $B_\sigma(\alpha)$ originate from the conversion of the momentum sum into an energy integral and read for parabolic dispersion:

$$B_\sigma(\alpha) = \frac{1}{\Gamma(\sigma)} \int_0^\infty \frac{dy y^{\sigma-1}}{e^{y+\alpha} - 1} = \sum_{\nu=1}^\infty \frac{e^{-\nu\alpha}}{\nu^\sigma} \quad [3]$$

with $B_\sigma(0) = \zeta(\sigma)$, Γ the Euler Γ -function and ζ denoting the Riemann ζ -function. It is important to understand that in order to have a constant total density n , $B_{3/2}(\alpha)$ has to increase $\propto T^{-3/2}$ in the same way as λ_T^3 . This is, however, impossible at all temperatures since the chemical potential of the Bose gas vanishes ($\alpha \rightarrow 0$) at a finite temperature T_B given by

$$T_B = \frac{2\pi\hbar^2}{mk_B} \left[\frac{n}{\zeta(3/2)} \right]^{2/3} \quad [4]$$

for which $n \lambda_{T_B}^3 = B_{3/2}(0) = \zeta(3/2) = 2.612 \dots$

In sharp contrast, fermions obey the Pauli exclusion principle, which states that only one fermion can occupy a quantum state $|k, \sigma\rangle$ specified in addition by the spin projection σ . The average statistical occupation is given by the Fermi-Dirac distribution

$$f_k = \frac{1}{e^{(\epsilon_k - \mu)/k_B T} + 1} \quad [5]$$

Figure 3 shows a comparison of Bose-Einstein and Fermi-Dirac momentum distributions n_k plotted vs. ϵ_k . The chemical potential is shown for fermions only, $\mu_F = k_B T \alpha$ is always positive and the total density can be expressed as

$$n = \frac{1}{V} \sum_{k, \sigma} f_k = \frac{2}{\lambda_T^3} F_{\frac{3}{2}}(\alpha) \quad [6]$$

where the factor of 2 originates from the spin degeneracy. For parabolic dispersion, the Fermi integral reads:

$$F_\sigma(\alpha) = \frac{1}{\Gamma(\sigma)} \int_0^\infty \frac{dy y^{\sigma-1}}{e^{y+\alpha} + 1} \stackrel{T \rightarrow 0}{=} \frac{(\mu/k_B T)^\sigma}{\Gamma(\sigma+1)} \quad [7]$$

One recognizes that the degeneracy condition $n \lambda_T^{3/2} \geq 1$ corresponds to the limit $T \leq T_F = \mu(0)/k_B$, which is connected with the formation of a “Fermi sea,” with $\mu(0) \equiv E_F$ the Fermi energy:

$$\mu \stackrel{T \rightarrow 0}{=} \frac{\hbar^2}{2m} (3\pi^2 n)^{2/3} = E_F \quad [8]$$

To summarize, quantum behavior in Bose and Fermi system sets in below the degeneracy temperature T^* , defined through $n \lambda_{T^*}^3 = O(1)$. For bosons, $T^* = T_B$ is the temperature at which the chemical potential vanishes, whereas for fermions $T^* = T_F$ is the Fermi temperature.

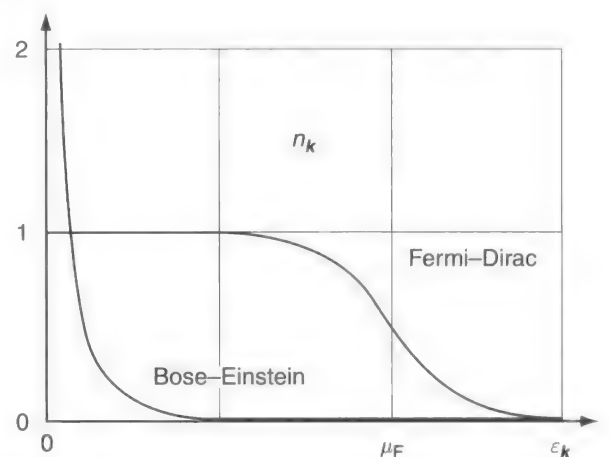


Figure 3 The Fermi and Bose momentum distribution.

London Quantum Hydrodynamics

For a general treatment of the quantum-mechanical origin of the equations describing Bose and Fermi superfluidity, it is convenient to introduce a parameter ν which describes single bosons ($\nu=1$) or Fermion pairs ($\nu=2$) of mass $M=\nu m$. The basic assumption (London 1938) is that the laws of quantum mechanics are applicable also to a macroscopic number of single ($\nu=1$) or composite ($\nu=2$) particles of density $\rho^s/\nu m$, the so-called condensate, which is represented by a macroscopic wave function $\psi(\mathbf{r}, t)$. ψ has the property

$$\psi(\mathbf{r}, t)\psi^*(\mathbf{r}, t) = \frac{\rho^s(\mathbf{r}, t)}{\nu m}; \quad \nu = 1, 2$$

The dynamics of the condensate is governed by the Schrödinger equation

$$i\frac{\hbar}{\nu} \frac{\partial \psi}{\partial t} = \left(-\frac{\hbar^2 \nabla^2}{2\nu^2 m} + \mu \right) \psi \quad [9]$$

in which μ represents the condensate's chemical potential. After performing a Madelung transformation (Madelung 1926):

$$\psi = ae^{i\varphi}, \quad a^2 = \frac{\rho^s}{\nu m}$$

one arrives at two coupled hydrodynamic equations, the first of which reads

$$\begin{aligned} \frac{\partial \rho^s}{\partial t} + \nabla \cdot \mathbf{j}_m^s &= 0 \\ \mathbf{j}_m^s &= \rho^s \mathbf{v}^s, \quad \mathbf{v}^s = \frac{\hbar}{\nu m} \nabla \varphi \end{aligned} \quad [10]$$

Equation [10] can be interpreted as a continuity equation, which represents the conservation law for the condensate mass density ρ^s . The second equation

$$-\frac{\hbar}{\nu} \frac{\partial \varphi}{\partial t} = \frac{1}{2} m \mathbf{v}^{s2} + \mu + O(\hbar^2 \nabla^2) \quad [11]$$

assumes the form of the Hamilton–Jacobi equation for the action field of classical mechanics $\hbar\varphi$, if the quasiclassical limit (terms $\propto O(\hbar^2 \nabla^2) \rightarrow 0$) is taken.

From [10] and [11] a condensate acceleration equation can be derived, which resembles the Euler equation of classical hydrodynamics ($\mu = \mu_0 + \delta\mu$):

$$\frac{\partial \mathbf{v}^s}{\partial t} + (\mathbf{v}^s \cdot \nabla) \mathbf{v}^s = -\frac{1}{m} \nabla \delta\mu \quad [12]$$

The physical nature of the driving force becomes evident after applying the Gibbs–Duhem relation

$n\delta\mu = \delta P - \sigma_0 \delta T$. Finally, the acceleration of the mass supercurrent \mathbf{j}_m^s is of the form

$$\frac{\partial \mathbf{j}_m^s}{\partial t} = -\frac{\rho^s}{\rho} \nabla (\delta P - \sigma_0 \delta T) \quad [13]$$

It turns out that the London equations [10] and [13], in which ρ^s is an unknown phenomenological parameter, explain many experimental observations such as persistent currents, U-tube oscillations, thermomechanical (e.g., fountain-) effects, beaker flow phenomena, and many others.

Bose–Einstein Condensation (BEC)

In order to understand the macroscopic quantum state in case of Bose systems, we consider first the simple case of a Bose gas. Let us decompose the energy eigenstates ϵ_k into those with $\epsilon_k = \epsilon_0 = 0$ (condensate) and average occupation number

$$n_0 = \frac{N_0}{V} = \frac{1}{V} \frac{1}{e^\alpha - 1} \quad [14]$$

and those with $\epsilon_k > 0$ (excited states) and average occupation number

$$n_{\text{ex}} = \frac{N_{\text{ex}}}{V} = \sum_k' n_k = n \frac{B_{3/2}(\alpha)}{B_{3/2}(0)} \left(\frac{T}{T_B} \right)^{3/2} \quad [15]$$

with the total density $n = n_{\text{ex}} + n_0$. The consequence of the chemical potential vanishing at T_B clearly is a macroscopic occupation of the ground state of the Bose gas:

$$N_0 \stackrel{\alpha \rightarrow 0}{\sim} \frac{1}{1 + \alpha + \dots - 1} = \frac{1}{\alpha} \rightarrow \infty \quad [16]$$

This phenomenon is referred to as BEC. Below T_B , $\alpha = 0$ and from [15] we see that

$$n_{\text{ex}} \stackrel{\text{free bosons}}{=} n \left(\frac{T}{T_B} \right)^{3/2}, \quad T < T_B \quad [17]$$

The average occupation of the ground state is given by

$$n_0(T) = n - n_{\text{ex}}(T), \quad T < T_B \quad [18]$$

It is important to understand that the number density of condensed particles n_{ex} has nothing to do with the current response function ρ^s (eqn [10]). A derivation of ρ^s will be given in the section “Local response of condensates and excitation gases.”

Let us now discuss the structure of the excitation spectrum, which will turn out to be crucial for the observability of superfluidity, in some more detail. Suppose that a macroscopic object of mass M moves through the superfluid. Then one may ask the question, at what velocity does this motion cause the creation of an excitation of energy E_p and momentum p . The condition can be formulated in terms of the velocity

difference $\mathbf{v}_i - \mathbf{v}_f$ as $E_p = M(\mathbf{v}_i^2 - \mathbf{v}_f^2)/2$ and $\mathbf{p} = M(\mathbf{v}_i - \mathbf{v}_f)$. Eliminating \mathbf{v}_f yields $\epsilon_p = \mathbf{p} \cdot \mathbf{v}_i + O(M^{-1})$ so that condition for the creation of an excitation leads to the so-called Landau critical velocity

$$v_L = \min \left\{ \frac{E_p}{|\mathbf{p}|} \right\} > 0 \quad [19]$$

It is immediately clear that for free bosons $v_L = 0$. This means that a free Bose gas can never be a superfluid, since drag forces on moving objects will start to act even at smallest velocities.

It turns out that interaction effects can drastically modify the nature of the elementary excitations. In 1947, Nikolai Bogoliubov showed (for the first time using the method of second quantization) that even in the limit of weak repulsive interactions the excitation spectrum is phonon-like $E_p = c|\mathbf{p}|$, with c the sound velocity. Lev Landau and Richard Feynman investigated the situation for superfluid ^4He , where the interactions between the atoms are far from weak. Landau (1947) postulated the following form for the excitation spectrum, for which Feynman (1953) gave the microscopic justification. At low momenta, the spectrum is phonon-like and linear in \mathbf{p} :

$$\lim_{p \rightarrow 0} E_p = E_p^{\text{phon}} = c|\mathbf{p}| \quad [20]$$

At higher momenta, the spectrum is reminiscent of that of crystal phonons in that E_p passes through a maximum, and then, at a characteristic momentum p_0 approaches the next minimum, which, however, is located at a finite energy Δ . Feynman called this part of the spectrum the “roton” (mass m_r) in an analogy with a “smoke ring,” since it is connected with the forward motion of a particle accompanied by a ring of back-flowing other particles:

$$\lim_{|\mathbf{p}| \rightarrow p_0} E_p = E_p^{\text{rot}} = \Delta + \frac{(|\mathbf{p}| - p_0)^2}{2m_r} \quad [21]$$

Figure 4 shows a sketch of the phonon-roton spectrum of superfluid ^4He . Clearly, the Landau

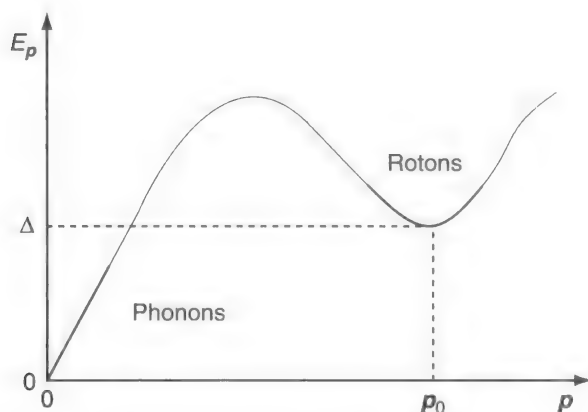


Figure 4 The phonon-roton spectrum.

critical velocity for the phonon-roton spectrum is characterized by the roton minimum and is given by $v_L \approx \Delta/p_0$.

BCS-Leggett Pair Condensation

The key assumptions of the weak-coupling mean-field BCS-Leggett pairing model can be summarized as follows: one first assumes that at sufficiently low temperatures it is energetically favorable that a temperature-dependent part of the fermions forms so-called Cooper pairs. This pair formation is caused by an attractive interaction in \mathbf{k} -space near the Fermi surface:

$$\Gamma_{kp}^{(s)} < 0, \quad |\xi_k|, |\xi_p| < \epsilon_c$$

Here $\xi_k = \epsilon_k - \mu$ measures the energy from the chemical potential. The index s denotes the total spin of the pair. Classical superconductors have pairs in a relative singlet state $s=0, m_s=0$ whereas the superfluid phases of liquid ^3He have pairs in a relative spin-triplet state $s=1, m_s=0, \pm 1$, with m_s the magnetic quantum number. The amplitude of spontaneous pair formation is

$$g_{k\sigma_1\sigma_2} \equiv \langle \hat{c}_{-k\sigma_1} \hat{c}_{k\sigma_2} \rangle \neq 0, \quad T \leq T_c \quad [22]$$

with $\mathbf{k} = \mathbf{k}_1 - \mathbf{k}_2$ the relative momentum of the pair. The attractive interaction that drives the Cooper-pair formation connects the pairing amplitude $g_{k\sigma_1\sigma_2}$ with a new energy scale, the so-called pair potential

$$\Delta_{k\sigma_1\sigma_2} = \sum_p \Gamma_{kp}^{(s)} g_{p\sigma_1\sigma_2} \quad [23]$$

As a consequence of triplet pairing the spin part of the pair potential is “even” upon interchange of σ_1 and σ_2 : $\Delta_{k\sigma_2\sigma_1} = \Delta_{k\sigma_1\sigma_2}$. Then the Pauli principle requires that $\Delta_{k\sigma_1\sigma_2}$ must be “odd” with respect to the interchange of \mathbf{k}_1 and \mathbf{k}_2 or, equivalently, $\mathbf{k} \rightarrow -\mathbf{k}$. The \mathbf{k} -dependence can now be classified by an orbital quantum number ℓ with the special cases of $\ell=1$ (p -wave) pairing, $\ell=3$ (f -wave) pairing, etc. All superfluid phases of ^3He are characterized by p -wave orbital symmetry.

The transition temperature T_c from [23] reads

$$k_B T_c = \frac{2e^\gamma}{\pi} \epsilon_c e^{-1/(N_F \Gamma^{(s)})}$$

with $N_F = 3n/2E_F$ the density of states at the Fermi level and $\gamma = 0.577\dots$ the Euler constant. The energies ξ_k can trivially be divided into particle-like ($\xi_k > 0$) and hole-like ($\xi_k < 0$) terms. The presence of the pair potential Δ_k leads to a mixing of particle- and hole-like contributions to the energy, which

becomes a matrix in particle-hole, or Nambu space (Nambu 1960), and generates what is referred to as off-diagonal long-range order (ODLRO):

$$\xi_k = \begin{pmatrix} \xi_k 1 & \Delta_k \\ \Delta_k^\dagger & -\xi_k 1 \end{pmatrix} \quad [24]$$

As usual, the diagonalization of ξ_k (Bogoliubov 1958) leads to the energy dispersion of the relevant thermal excitations of the superfluid state, the so-called Bogoliubov quasiparticles or “bogolons”:

$$E_k = \sqrt{\xi_k^2 + \Delta_k^2}, \quad \Delta_k^2 = \Delta_k \cdot \Delta_k^\dagger \quad [25]$$

In Figure 5, the dispersion E_p of Bogoliubov quasiparticles vs. $|p|$ is shown. It turns out that the superfluid phases (A and B) of liquid ^3He in zero magnetic field are characterized by unitary matrices Δ_k , so that the scalar quantity Δ_k can be interpreted as the energy gap in the bogolon spectrum, which, in general, may be anisotropic in k -space.

The energy gap Δ_k of the superfluid B-phase can be represented in the simple nodeless (pseudoisotropic) and BCS-like form (Balian and Werthamer, (BW), 1963):

$$\Delta_k = \Delta(T), \quad \frac{\Delta(0)}{k_B T_c} = \frac{\pi}{e^\gamma} \quad [26]$$

Its spin structure is characterized by the presence of all three triplet components $m_s = 0, \pm 1$ and will be discussed further with respect to the magnetization response (see next section). The gap symmetry of ^3He -A is uniaxial with respect to an axis $\hat{\ell}$ (Anderson and Morel 1960; Anderson and Brinkman 1973)

$$\Delta_k = \Delta_0(T) \sin \phi_k, \quad \frac{\Delta_0(0)}{k_B T_c} = \frac{\pi e^{5/6}}{2e^\gamma} \quad [27]$$

where $\cos \phi_k = \mathbf{k} \cdot \hat{\ell}$, and characterized by two point nodes of Δ_k at the zeros ($\phi_k = 0, \pi$) on the Fermi surface. It has furthermore turned out that only the

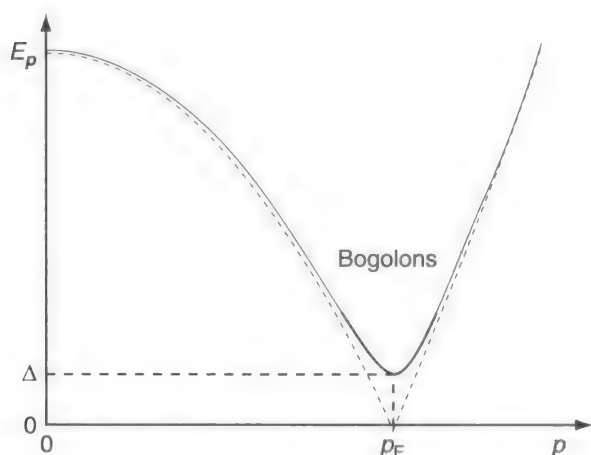


Figure 5 The bogolon energy dispersion.

$m_s = \pm 1$ components of the spin triplet contribute to its spin dependence (equal spin pairing (ESP)).

Local Response of Condensates and Excitation Gases

In the previous sections we have seen that the structure (energy dispersion, statistics, critical flow velocity) of the relevant thermal excitations is of crucial importance for the superfluidity. We can now aim at a generalized statistical description of bosonic (phonons, rotons) and fermionic (bogolons) excitation gases, by introducing a generalized momentum distribution

$$n_\theta\{E_k\} = \frac{1}{e^{E_k/k_B T} - \theta} \quad [28]$$

and its energy derivative

$$\varphi_{k,\theta} = -\frac{\partial n_\theta\{E_k\}}{\partial E_k} = \frac{1}{2k_B T [\cosh(E_k/k_B T) - \theta]} \quad [29]$$

Special cases are

$$\theta = \begin{cases} 1, & \text{Bose (phonons, rotons)} \\ -1, & \text{Fermi (bogolons)} \end{cases}$$

Introducing the spin $s = (1 - \theta)/4$, the total momentum density response to the presence of a superfluid velocity

$$\mathbf{v}^s = \frac{\hbar \nabla \varphi}{(2s + 1)m}, \quad s = 0, \frac{1}{2}$$

and a normal fluid velocity \mathbf{v}^n can be written in the general form

$$\mathbf{j}_m = \frac{2s + 1}{V} \sum_k p n_\theta\{E_k + \delta E_k\} + \rho \mathbf{v}^s \quad [30]$$

After Taylor-expanding n_θ with respect to the small energy shifts $\delta E_k = \mathbf{p} \cdot (\mathbf{v}^s - \mathbf{v}^n)$, one may introduce the so-called normal fluid density tensor

$$\rho_{ij}^n = \frac{2s + 1}{V} \sum_k \varphi_{k,\theta} p_i p_j \quad [31]$$

and the momentum density assumes the form

$$\mathbf{j}_m = \rho^s \cdot \mathbf{v}^s + \rho^n \cdot \mathbf{v}^n, \quad \rho^s = \rho 1 - \rho^n \quad [32]$$

Equation [32] forms the central result of this essay because it represents the microscopic counterpart of the generalized London equation [10]. It is clearly seen how the phenomenon of superfluidity originates from $\rho^s > 0$ due to a qualitative change in the dispersion of the elementary excitations, which may in particular be characterized by a gap in

the excitation spectrum. Equation [32] is more general than [10] in that it introduces a two-fluid picture in which the mass supercurrent $\mathbf{j}_m^s = \rho^s \mathbf{v}^s$ (eqn [10]) is complemented by a normal (excitation) mass current $\mathbf{j}_m^n = \rho^n \mathbf{v}^n$ in the presence of a macroscopic velocity field \mathbf{v}^n of the excitation gas obeying arbitrary statistics. The temperature dependence of $\rho^s(T)$ can now be computed via [31] and the result depends on the dispersion of the thermal excitation under consideration. Figure 6 shows the temperature dependence of the normal and superfluid density of superfluid ^4He . The normal fluid density of superfluid ^3He is, in general, a tensor quantity

$$\rho_{ij}^n = \begin{cases} \rho_{\parallel}^n \hat{\ell}_i \hat{\ell}_j + \rho_{\perp}^n (\delta_{ij} - \hat{\ell}_i \hat{\ell}_j), & ^3\text{He-A} \\ \rho^n \delta_{ij}, & ^3\text{He-B} \end{cases} \quad [33]$$

The short-range Fermi liquid interaction leads to a quasiparticle mass enhancement $m^*/m = 1 + F_1^s/3$ characterized by the pressure-dependent dimensionless Landau parameter F_1^s . In Figure 7, the normal fluid density ($\rho_{\parallel,\perp}^n$ for $^3\text{He-A}$, ρ^n for $^3\text{He-B}$) is shown as a function of reduced temperature at a pressure of 27 bar, where $F_1^s = 12.53$. The entropy density of an excitation system of arbitrary statistics below the transition can be written as

$$\sigma_0 = k_B \frac{(2s+1)}{V} \sum_k P_{k\theta} \quad [34]$$

$$P_{k\theta} = \theta(1 + \theta n_{\theta}) \ln(1 + \theta n_{\theta}) - n_{\theta} \ln n_{\theta}$$

with $n_{\theta} = n_{\theta}(E_k)$, from which one may derive the specific heat capacity

$$\begin{aligned} T\delta\sigma &= \frac{2s+1}{V} \sum_k E_k n_{\theta} \left\{ \frac{E_k + \frac{\partial E_k}{\partial T} \delta T}{k_B(T + \delta T)} \right\} \\ &= c_V \delta T \end{aligned} \quad [35]$$

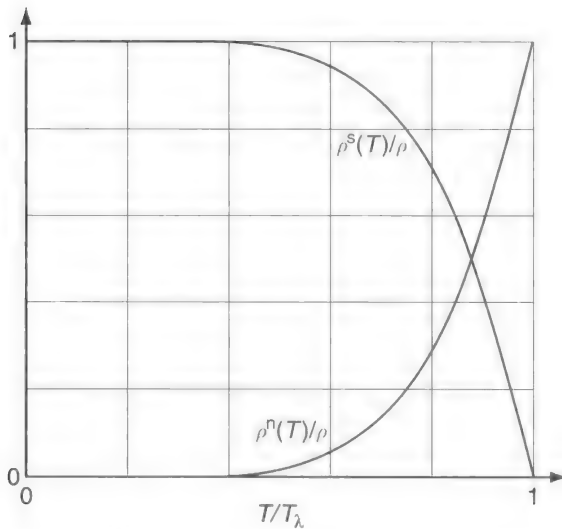


Figure 6 The normal and superfluid density for He-II.

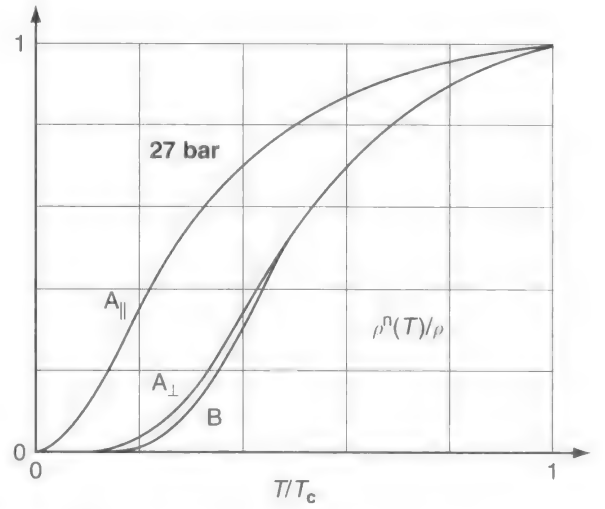


Figure 7 The normal fluid density for $^3\text{He-A, B}$.

After a Taylor expansion of n_{θ} with respect to the small local temperature change δT , the result for $c_V(T)$ reads

$$c_V = \frac{2s+1}{V} \sum_k \varphi_{k,\theta} \left(\frac{E_k^2}{T} - E_k \frac{\partial E_k}{\partial T} \right) \quad [36]$$

In Figures 8 and 9 we show the cusp-like specific heat of a Bose gas as compared with the specific heat of $^3\text{He-A, B}$, which display discontinuities at T_c .

Finally, the superfluid phases of ^3He are characterized in addition by the spin degrees of freedom, reflected by the bogolon spin magnetization response to an external magnetic field B :

$$M^n = \frac{\gamma \hbar}{2V} \sum_{k,\sigma} \sigma n_{-1} \{E_k - \gamma \hbar \sigma B/2\} = \chi_0 B \quad [37]$$

where γ denotes the gyromagnetic ratio of the fermions. The bogolon spin susceptibility χ_0 is obtained after a Taylor expansion of n_{-1} with respect to B as

$$\chi_0 = \left(\frac{\gamma \hbar}{2} \right)^2 \frac{1}{V} \sum_{k,\sigma} \varphi_{k,-1} \equiv \left(\frac{\gamma \hbar}{2} \right)^2 N_F Y(T) \quad [38]$$

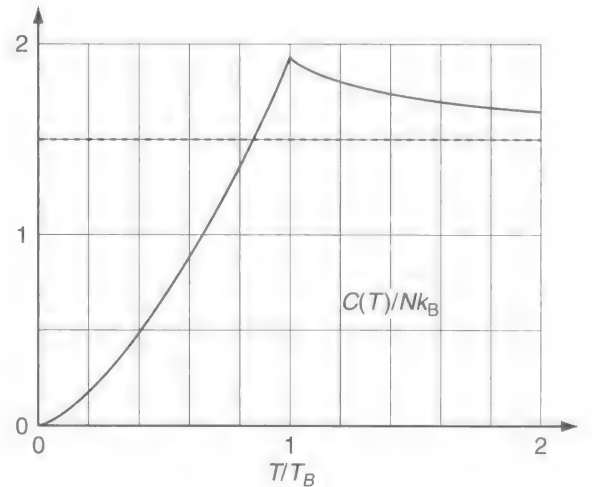


Figure 8 The specific heat capacity of a Bose gas.

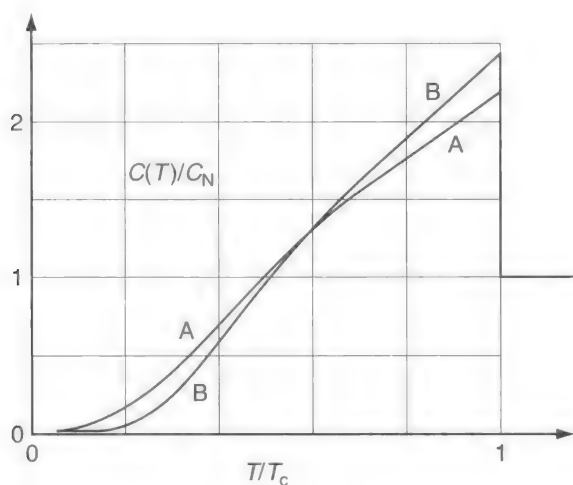


Figure 9 The specific heat of ³He-A, B.

Note that eqn [38] accounts only for the $m_s = 0$ (bogolon) contribution to the spin-triplet susceptibility, the temperature dependence of which is given by the so-called Yosida function $Y(T) = N_F^{-1} \sum_{k\sigma} \varphi_{k,-1}$. The total susceptibility reads

$$\chi_{\text{tot}} = \underbrace{\chi_0}_{\text{bogolons}} + \underbrace{\chi_1 + \chi_{-1}}_{\text{condensate}} \tag{39}$$

with the condensate contributing through $\chi_{m_s=\pm 1}$ a fraction of 2/3 of the normal state Pauli susceptibility. In Figure 10, the reduced spin susceptibility χ/χ_N of ³He-A, B is plotted vs. reduced temperature. While the constant susceptibility is characteristic of the ESP pairing state, the reduction of the B-phase susceptibility is due to the lack of the nonmagnetic $m_s = 0$ contribution to the spin triplet in the low-temperature limit. Exchange interaction effects, characterized by the dimensionless Landau parameter F_0^a , lead to a further reduction of the Balian-Werthamer (BW)-state susceptibility, which is shown for 27 bar, where $F_0^a = -0.755$. Note that the theoretical picture reflected in Figure 10, and also in Figures 6, 7, and 9, is in quantitative agreement with experimental observations.

In summary, superfluidity is a quantum-mechanical phenomenon seen on a macroscopic scale. It occurs below the degeneracy temperature $T^* \propto n^{2/3}/m$ of both Bose and Fermi many-particle systems (like liquid ⁴He and ³He) and is a property of a macroscopic number of particles, the condensate. The role of (weak or strong) interactions is manifested in the structure of the relevant elementary excitations, which always exist in addition to the condensate at finite temperatures and above certain critical velocities. These excitations form a gas, referred to as the normal fluid, since it gives rise to temperature-dependent thermodynamic and response

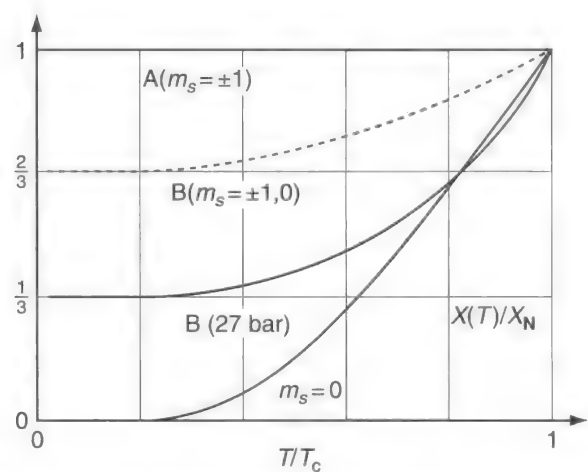


Figure 10 The spin susceptibility of ³He-A, B.

functions and contributes to the entropy and the flow dissipation. Superfluidity is now well understood using various aspects of the concept of the macroscopic wave function. On the microscopic level, the mechanisms of BEC and BCS–Leggett pair formation have been successfully invoked to understand the fascinating properties of Bose and Fermi superfluids.

See also: Bose–Einstein Condensates; Bosons and Fermions in External Fields; High T_c Superconductor Theory; Topological Knot Theory and Macroscopic Physics; Variational Techniques for Ginzburg–Landau Energies; Vortex Dynamics.

Further Reading

Anderson PW and Brinkman WF (1973) *Physical Review Letters* 30: 1911.
Anderson PW and Morel P (1960) *Physical Review Letters* 5: 136.
Annett J (2004) *Superconductivity, Superfluids and Condensates*. Oxford: Oxford University Press.
Balian R and Werthamer NR (1963) *Physical Review* 131: 1553.
Bogoliubov NN (1947) *Journal of Physics USSR* 11: 23.
Bogoliubov NN (1958) *Soviet Physics JETP* 7: 794.
Dobbs ER (2000) *Helium Three*. Oxford: Oxford University Press.
Feynman RP (1953) *Physical Review* 91: 1301.
Keller WE (1969) *Helium-3 and Helium-4*. New York: Plenum.
Landau LD (1947) *Journal of Physics USSR* 11: 91.
London F (1938) *Physical Review* 54: 947.
Madelung E (1926) *Z. Physik* 40: 322.
Nambu Y (1960) *Physical Review* 117: 648.
Tilley DR and Tilley J (1990) *Superfluidity and Superconductivity*. Adam Hilger.
Tisza L (1938) *Journal de Physique et Radium* 1: 164.
Tsuneto T (1998) *Superconductivity and Superfluidity*. Cambridge: Cambridge University Press.
Vollhardt D and Wölfle P (1990) *The Superfluid Phases of Helium 3*. Taylor and Francis.

Supergravity

K S Stelle, Imperial College, London, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction: Minimal $D = 4$ Supergravity

The essential idea of supersymmetry is an extension of the relativistic structure group of spacetime, which in ordinary four-dimensional physics in the absence of gravity is the Poincaré group $ISO(3, 1)$. In a minimal supersymmetric theory in flat $D = 4$ spacetime, the minimal supersymmetry algebra (the “graded Poincaré algebra”) adds spinorial generators Q_α to the Lorentz generators M_{mn} and the translational generators (momenta) P_m , where $m = 0, 1, 2, 3$. The core relation is the “anticommutator” of two Q_α :

$$\{Q_\alpha, \bar{Q}_\beta\} = -2\gamma_{\alpha\beta}^m P_m \quad [1]$$

where $\bar{Q} = Q^\dagger \gamma^0$ and the γ^m are the Dirac gamma matrices. In the minimal $D = 4$ supersymmetry algebra, the spinor generator Q_α is taken to be Majorana: $Q = C(\bar{Q})^T$, where C is the charge-conjugation matrix and A^T denotes the transpose of the matrix A . The full supersymmetry algebra adjoins to the anticommutation relation [1] the usual commutation relations among the Lorentz generators and the commutators of the Lorentz generators with the momenta and the spinors Q_α ; the latter express respectively the vectorial and spinorial characters of P_m and Q_α :

$$i[M_{mn}, M_{pq}] = \eta_{np} M_{mq} - \eta_{mp} M_{nq} \quad [2]$$

$$i[M_{mn}, P_q] = \eta_{nq} P_m - \eta_{mq} P_n \quad [3]$$

$$i[M_{mn}, Q_\alpha] = \frac{1}{2}(\gamma_{mn} Q)_\alpha \quad [4]$$

where $\gamma_{mn} = (1/2)(\gamma_m \gamma_n - \gamma_n \gamma_m)$ and $\eta_{mn} = \text{diag}(-1, 1, 1, 1)$ is the Minkowski metric. The final relation in the supersymmetry algebra expresses the flatness of Minkowski space:

$$[P_m, P_n] = 0 \quad [5]$$

This algebra has been considered as an extension of the symmetry algebra of particle physics since the work of Gol’fand and Likhtman in 1971, and especially since the linearly realized supersymmetric model of Wess and Zumino in 1974. That model contains a pair of $D = 4$ scalar fields and a $D = 4$ Majorana spinor, so the numbers of bosonic and fermionic degrees of freedom are equal; this is a fundamental characteristic of supersymmetric theories.

The work of Wess and Zumino led to an explosion of interest in supersymmetry, especially

once it was realized that renormalizable supersymmetric models display a cancellation of some of the divergences that have plagued relativistic quantum field theory since its inception in the 1930s. In particular, in renormalizable flat-space field theory models, divergences quadratic in a high-momentum cutoff vanish as a result of cancellations between virtual bosonic and fermionic particles. This is a very attractive feature for control of the “hierarchy problem” in particle physics, especially for the instability inherent in having vastly different scales within the same theory, for example, the TeV scale of ordinary electroweak physics and the 10^{16} GeV scale where unification with the strong interactions might come in.

When one includes gravity, the stability problems of particle physics become much more severe. Einstein’s theory of general relativity is itself non-renormalizable, that is, its ultraviolet divergences are of different forms from the terms present in the original “classical” action and there is no acceptable finite set of correction terms that can be added to it to remove this defect. Moreover, when otherwise tolerably behaved matter field theories that are renormalizable in a flat-spacetime context are coupled to general relativity, the gravitational couplings pollute the matter theories with non-renormalizable divergences. This is a key aspect of the great difficulty that has been encountered in interpreting gravity as a quantum theory.

Supersymmetry, with its divergence-canceling powers, was thus a very attractive option in the struggle to formulate a quantum theory of gravity, and the creation of a supergravity theory was thus a very high priority task. This was achieved in 1976 by Freedman, Ferrara, and Van Nieuwenhuizen using the technique of iterative Noether coupling to build up this nonlinear theory order-by-order in powers of the fermionic fields. The fermionic partner of the massless spin-2 “graviton” field is a massless fermionic spin-3/2 field that has come to be called the “gravitino.”

A second 1976 paper by Deser and Zumino soon followed, emphasizing how supergravity manages to circumvent the well-known problems of coupling spins higher than 1 to gravity. A key point in achieving this result is the role played by the local version of the supersymmetry algebra [1]–[5]. As one can see from the translations occurring on the right-hand side of [1], when one replaces translation symmetry by local general coordinate invariance in a gravitational context, the supersymmetry transformations must themselves become local as well. Local symmetries allow for transformation parameters

that are local in the spacetime coordinates x^m , and in interacting theories they require coupling of the corresponding “gauge field” to a conserved current. In the case of supergravity, the gravitino field $\psi_{m\alpha}$ plays this gauge-field role, and its coupling to the conserved current of supersymmetry is the key to allowing a consistent coupling between the spin-2 graviton and the spin-3/2 gravitino.

The Minimal Supergravity Action

The action for minimal supergravity in $D=4$ dimensions can be written, using the vierbein formalism where the metric is expressed as a quadratic expression in a nonsymmetric 4×4 vierbein matrix e_m^a , $g_{mn} = e_m^a e_n^b \eta_{ab}$, as

$$I = \frac{1}{2\kappa^2} \int d^4x \det(e) R(e, \omega(e) + K(\psi)) - \frac{i}{2} \int d^4x \epsilon^{mnpq} \bar{\psi}_m \gamma_5 \gamma_n D_p(e, \omega(e) + K(\psi)) \psi_q \quad [6]$$

where $\kappa = \sqrt{8\pi G}$ is the gravitational coupling constant,

$$\omega_m^{ab}(e) = \frac{1}{2} e^{na} (e_{n,m}^b - e_{m,n}^b) - \frac{1}{2} e^{nb} (e_{n,m}^a - e_{m,n}^a) + \frac{1}{2} e^{na} e^{rb} (e_{nc,r} - e_{rc,n}) e_m^c \quad [7]$$

is the usual vierbein formalism spin connection (in which $e_{n,m}^b = \partial_m e_n^b$ and e^{ma} is the matrix inverse of e_{ma}), and

$$K_m^{ab}(\psi) = \frac{i\kappa^2}{4} (\bar{\psi}_m \gamma^a \psi^b + \bar{\psi}^a \gamma_m \psi^b - \bar{\psi}_m \gamma^b \psi^a) \quad [8]$$

is the fermionic contorsion, an additional part of the covariant derivative $D_m(e + K(\psi))$ appearing in the action [6]. (Indices m, n are taken to be “world” indices while indices a, b are “tangent space” indices; one can convert from one type to another using the vierbein e_m^a and its inverse, e.g., $\psi_{a\alpha} = e_m^a \psi_{m\alpha}$.)

Keeping the terms in the action grouped as above using the nonstandard covariant derivative $e_m^{ab} + K_m^{ab}$ is what has been called “1.5 order formalism”: this greatly simplifies the writing and analysis of the supergravity action [6]. In the action [6], one has the Ricci scalar $R(e, \omega(e) + K(\psi))$ written in terms of this generalized torsional spin connection. One may of course expand out all the $\omega_m^{ab} + K_m^{ab}$ combinations and write the nonlinear fermionic terms separately. Doing this produces a quartic term

$$L_4 = \frac{\kappa^2}{32} [\bar{\psi}^b \gamma^a \psi^c (\bar{\psi}_b \gamma_a \psi_c + 2\bar{\psi}_a \gamma_b \psi_c) - 4(\bar{\psi}^a \gamma^b \psi^c)(\bar{\psi}_a \gamma_b \psi_c)]$$

showing the highly nonlinear nature of supergravity theory – when expanded out, the theory becomes much more cumbersome to study. The 1.5 order formalism trick is one of a large number of algebraic simplifications that had to be developed in order to master the technical aspects of supergravity. It also reveals a characteristic physical feature: this theory naturally involves a connection with torsion built from the fermionic fields.

In terms of the torsional covariant derivative $D_m \epsilon(x) = (\partial_m + (1/4)(\omega_m^{ab}(e) + K_m^{ab}(\psi))\gamma_{ab})\epsilon(x)$ of the infinitesimal supersymmetry parameter $\epsilon(x)$, the local supersymmetry transformations which leave the action [6] invariant (up to the integral of a total derivative) are

$$\delta e_m^a = i\bar{\epsilon} \gamma_a \psi_m \quad [9]$$

$$\delta \psi_m = 2\kappa^{-1} D_m \epsilon \quad [10]$$

The inhomogeneous part $2\kappa^{-1} D_m \epsilon$ in the gravitino transformation [10] demonstrates the gauge-field nature of the gravitino field. For a distribution of “supermatter” fields (e.g., Wess–Zumino model scalars and spinors), the integrated “charge” that one would get from a Gauss’s law surface integral at spatial infinity using the gravitino gauge field is the total supercharge Q_α , which in turn plays the role of the supersymmetry generator in the original matter-sector supersymmetry algebra [1].

Both the gravitational field and the gravitino field are thus effectively gauge fields, albeit not of a standard Yang–Mills type. The local algebra is a deformation of the rigid supersymmetry algebra [1]–[5], generalizing the relation between general covariance and flat-space Poincaré symmetry. Some basic consequences of the flat-space algebra are preserved, however. An extremely important instance of this is energy positivity. As one can see by multiplying [1] by γ^0 and then contracting on the spinor index,

$$E = P^0 = \frac{1}{2} \sum_\alpha \{Q_\alpha, Q_\alpha^\dagger\}$$

The right-hand side is manifestly non-negative provided the theory is quantized in a positive-metric Hilbert space. One can see this even more explicitly in a Majorana spinor basis, where $Q_\alpha^\dagger = Q_\alpha$. Accordingly, for flat-space supersymmetric theories, one obtains directly the result that energy is non-negative. This carries over to the local algebra of supergravity, where the total energy is obtained from a Gauss’s law integral over the sphere at spatial infinity.

In general relativity, an integrated energy can be defined with respect to an asymptotic timelike Killing vector at spatial infinity. Showing that this

energy is non-negative remained for decades a famously unsolved problem in gravitational physics; it was ultimately proven in Yau's positive-energy theorem. The algebraic structure of supergravity makes energy positivity much more transparent, however. Since pure general relativity can be obtained by setting the gravitino field to zero, this result is inherited by pure Einstein theory as a consequence of its being embeddable into supergravity. Energy positivity can thus be proved even at the classical level using ideas taken from supergravity, as was done by Witten and later streamlined by Nester, in an argument much simpler than Yau's proof. This argument writes the energy as an integral over a positive-semidefinite expression quadratic in a commuting spinor field which is analogous to the (anticommuting) spinor parameter of supergravity in the transformations [9] and [10].

Auxiliary Fields and Superspace

Supergravity shares with flat-space supersymmetric theories a curious technical feature that gives a hint of a new underlying geometry. Standard counting of the gauge-invariant continuous degrees of freedom of the graviton and the gravitino in momentum space yield the same result per momentum value: two bosonic degrees of freedom and two fermionic degrees of freedom. This accords with the general requirement in supersymmetric theories that the numbers of bosonic and fermionic degrees of freedom match. This count follows from the Einstein and spin-3/2 equations of motion, or "on-shell." If one compares the count of nongauge degrees of freedom without using the equations of motion (i.e., "off-shell"), one obtains an imbalance, however: six nongauge graviton versus 12 nongauge fermion fields. This is directly related to another puzzling feature of the supergravity realization of local supersymmetry: the local supersymmetry algebra closes onto a finite set of transformations only when the equations of motion are imposed.

As in flat-space supersymmetry, the cure for this problem is to add nondynamical "auxiliary" fields to the action. In the supergravity case, the imbalance in the off-shell bose-fermi field count indicates that an additional six bosonic fields are needed. In the minimal set of auxiliary fields, these organize into a vector b_m and a scalar-pseudoscalar pair M, N ; the additional terms in the action [6] are simply

$$\int d^4x \det(e) \left(-\frac{1}{3} M^2 - \frac{1}{3} N^2 + \frac{1}{3} b_m b^m \right)$$

while the local supersymmetry transformations are changed to include the auxiliary fields, e.g., the gravitino transformation becomes

$$\delta\psi_m = 2\kappa^{-1} D_m(\omega, K)\epsilon + \gamma_5 \left(b_m - \frac{1}{3} \gamma_m \gamma^n b_n \right) \epsilon - \frac{1}{3} \gamma_m (M + \gamma_5 N) \epsilon$$

while the auxiliary fields transform into expressions that vanish on-shell. Since the field equations for the auxiliary fields are algebraic in character and since for source-free supergravity they have the simple solution $b_m = M = N = 0$, one can directly regain the on-shell formalism by algebraically eliminating the auxiliary fields.

The inclusion of auxiliary fields is not an empty trick, however. The local supersymmetry transformations including the auxiliary fields form a closed set without the use of equations of motion ("off-shell closure"). This standardizes the form of the supersymmetry transformations so that they remain the same even when supermatter is coupled to supergravity instead of needing a case-by-case Noether construction as in the case without the auxiliary fields. In this way, a standard set of coupling rules can be drawn up, known as the "tensor calculus." This tensor calculus is of great importance as it allows for the construction of general models of supergravity coupled to supermatter (Wess–Zumino multiplets and super Yang–Mills multiplets consisting of spin-1 gauge fields and spin-1/2 "gaugino" fields). These general couplings form the basis for essentially all supersymmetric phenomenology, and in particular for the formulation of the Minimal Supersymmetric Standard Model. Since supersymmetry is not directly observed in low-energy physics, it must be spontaneously broken, like many other gauge symmetries. As it happens, the physically realistic mechanisms of supersymmetry breaking all originate from supergravity couplings derived using the tensor calculus.

Given the regular set of tensor calculus rules for coupling supergravity to supermatter, one is led to suspect that a geometrical structure lies in the background. This is indeed the case; the corresponding construction is known as "superspace."

The basic idea of superspace is a generalization of the coset space construction of Minkowski space as the coset space given by the Poincaré group divided by the Lorentz group: $\mathcal{M}_4(x^m) = \text{ISO}(3, 1)/\text{SO}(3, 1)$. For supersymmetric theories, one analogously constructs $\text{Superspace}(x^m, \theta^\alpha) = \text{Graded Poincaré}/\text{SO}(3, 1)$. The basic ideas of superspace were introduced by Akulov and Volkov in 1972, while the idea of expanding in "functions" on this space, thus yielding "superfield," was introduced by Salam and Strathdee

in 1974. This led to a formulation of the Wess–Zumino model in terms of a chiral superfield $\phi(x, \theta)$, which is subjected to a covariant superspace constraint.

In order to manage the formalism of superspace more efficiently, it is convenient to use a two-component spinor formalism corresponding to the Weyl basis for the Dirac gamma matrices, in which the Majorana spinor coordinate θ is represented as

$$\theta = \begin{pmatrix} \theta_\alpha \\ \bar{\theta}^{\dot{\alpha}} \end{pmatrix}$$

where two-component indices $\alpha, \dot{\alpha} = 1, 2$ are raised and lowered with the covariant two-index antisymmetric tensors $\epsilon^{\alpha\beta}, \epsilon^{\dot{\alpha}\dot{\beta}}$, which both take the numerical value $i\sigma_2$. The flat-space fermionic covariant derivatives are then

$$\begin{aligned} D_\alpha &= \frac{\partial}{\partial \theta^\alpha} + i\sigma_{\alpha\dot{\beta}}^m \bar{\theta}^{\dot{\beta}} \partial_m \\ \bar{D}_{\dot{\alpha}} &= -\frac{\partial}{\partial \bar{\theta}^{\dot{\alpha}}} + i\bar{\theta}^{\beta} \sigma_{\beta\dot{\alpha}}^m \partial_m \end{aligned} \quad [11]$$

where the $\sigma_{\alpha\dot{\beta}}^m = (1, \sigma_i)$ for $m = (0, i)$ (where σ_i are the Pauli matrices) are the Van der Waerden matrices which establish the mapping between vector indices and (chiral, antichiral) spinor index pairs. The Wess–Zumino multiplet is then described by a complex chiral superfield satisfying the constraint $\bar{D}_{\dot{\alpha}}\phi = 0$. Unlike the situation in Minkowski space, where the only Lorentz-covariant solution to a constraint that sets to zero the $\partial/\partial x^m$ derivatives is a constant, superspace has a reducible set of coordinates $(x^m, \theta^\alpha, \bar{\theta}^{\dot{\alpha}})$ and, as a result, requiring ϕ to be annihilated by $\bar{D}_{\dot{\alpha}}$ does not require the whole superfield to be a constant.

Since the fermionic coordinates of superspace $\theta^\alpha, \bar{\theta}^{\dot{\alpha}}$ are anticommuting (i.e., they are elements of a Grassman algebra), and since $\alpha, \dot{\alpha} = 1, 2$ have an index range of two, powers of them higher than the second order necessarily vanish. As a result, superfields like ϕ can be expanded into sets of component fields, each of which is an ordinary field in Minkowski space. In this way, a chiral superfield expands into $(A(x), B(x), \chi_\alpha(x), \bar{\chi}_{\dot{\alpha}}(x), F(x), G(x))$, where the fields A, B, χ , and $\bar{\chi}$ are the physical fields of the Wess–Zumino model, while F and G are dimension-2 auxiliary fields. In this way, the auxiliary fields of supersymmetry naturally fit into a superspace formalism as higher components in a superfield expansion. It is in this sense that they point toward the superspace formulations of supersymmetric theories.

For supergravity, there are a number of different approaches to realizing the theory in superspace,

and these correspond naturally to the various possible choices of auxiliary-field sets. With the minimal set, the supergravity multiplet is described by a superfield carrying a vector index $H_m(x, \theta, \bar{\theta})$; this superfield is called the prepotential of supergravity. Note the fact that since the divisor group in the coset-space construction of superspace is the Lorentz group, superfields may carry indices corresponding to any Lorentz representation. The component-field expansion of the H_m superfield yields the physical $e_m^a, \psi_{m\alpha}, \bar{\psi}_{m\dot{\alpha}}$ and auxiliary fields (b_m, M, N) together with a number of other components of dimension lower than those of the physical fields. This is not, however, all that surprising: even the physical fields $e_m^a, \psi_{m\alpha}, \bar{\psi}_{m\dot{\alpha}}$ contain components that are not directly related to the physical modes because we are dealing with a gauge theory. What occurs in superspace is a redundant expression of the supergravity multiplet with the presence of various component gauge fields.

The full expression of local supersymmetry in superspace can be given in a number of different formalisms. Suffice it here to indicate the transformation of the linearized theory expanded in small fluctuations about empty flat superspace. Converting the vector index of H_m into a (chiral, antichiral) spinor index pair via $H_{\alpha,\dot{\beta}} = \sigma_{\alpha\dot{\beta}}^m H_m$, the linearized local symmetry transformation of the supergravity multiplet is

$$\delta H_{\alpha\dot{\beta}} = D_\alpha \bar{L}_{\dot{\beta}} - \bar{D}_{\dot{\beta}} L_\alpha \quad [12]$$

where the transformation parameter superfield L_α carrying a spinor index is antichiral: $D_\alpha \bar{L}_{\dot{\beta}} = 0$ (while the conjugate parameter superfield $\bar{L}_{\dot{\alpha}}$ is chiral). Expanding in component fields and comparing with the expansion of H_m , one sees that the chiral spinor superfield contains precisely the components needed to provide the standard gauge symmetries of e_m^a and $\psi_{m\alpha}, \bar{\psi}_{m\dot{\alpha}}$ and also to transform the other gauge components of H_m as well. One can then make various gauge choices according to taste in a given context.

One frequently encountered superspace gauge choice sets to zero all the fields in H_m except for the physical and auxiliary fields $(e_m^a, \psi_{m\alpha}, \bar{\psi}_{m\dot{\alpha}}, b_m, M, N)$. This is called a Wess–Zumino gauge following the analogy to a similar construction for super Maxwell theory (containing spins 1 and 1/2). Wess–Zumino gauge choices are not, however, supersymmetrically covariant. This shows up when one works out the supersymmetry algebra in such a gauge: the presence of auxiliary fields gives closure, as required, without use of the equations of motion, but the anticommutator of two supersymmetry

transformations when acting on a gauge field such as the Maxwell field or the vierbein gives a combination of the anticipated translation with an admixture of a gauge transformation with a field-dependent parameter.

The prepotential superfield of minimal supergravity can itself be fit into larger formalisms in superspace that are analogous to standard differential geometry, with supervielbeins, superspin connections and so forth. An unavoidable feature of these more seemingly geometric constructions, however, is their high degree of redundancy: superspace vielbeins and spin connections carrying Lorentz indices have many component fields in addition to those found in the prepotential. This redundancy is then cut down in turn by imposing superspace constraints on the geometrical superfields, for example, on the components of the torsion tensor in superspace.

Extended Supergravities and Supergravities in Higher Dimensions

The possible graded extensions of the Poincaré algebra allow for more than one spinorial generator. Thus, one can have N supersymmetry generators $Q_\alpha^i, \bar{Q}_{\dot{\beta}j}, i, j = 1, \dots, N$, with basic anticommutators (in Lorentz two-component notation)

{Q_\alpha^i, \bar{Q}_{\dot{\beta}j}} = 2\delta_j^i \sigma_{\alpha\dot{\beta}}^m P_m [13]

{Q_\alpha^i, Q_\beta^j} = 2\epsilon_{\alpha\beta} a^{\ell ij} Z_\ell [14]

{\bar{Q}_{\dot{\alpha}i}, \bar{Q}_{\dot{\beta}j}} = 2\epsilon_{\dot{\alpha}\dot{\beta}} \bar{a}^{\ell ij} Z_\ell [15]

The right-hand sides of [14] and [15] allow for the possibility of nonvanishing commutators between supersymmetry generators of the same chirality. As one can see from the overall symmetry in pairs of indices $(\alpha i, \beta j)$, the coefficients $a^{\ell ij}$ must be antisymmetric in the i, j indices, so such nonvanishing same-chirality anticommutators cannot occur for $N=1$. The corresponding abelian generators Z_ℓ are called central charges since they must commute with all the other $(Q_\alpha^i, \bar{Q}_{\dot{\beta}j}, P_m)$ elements of the algebra.

The i, j indices may be endowed with a symmetry meaning as well, although this is not obligatory in every model. When the central charges are absent, $Z_\ell=0$, one has $U(N)$ (or $SU(N)$) as the maximal such external automorphism; the choice of index placement on Q_α^i and $\bar{Q}_{\dot{\beta}j}$ anticipates this. If such a symmetry is realized in a given model, the fact that the $Q_\alpha^i, \bar{Q}_{\dot{\beta}j}$ carry representations both for that symmetry and for the spacetime Poincaré symmetry demonstrates how supersymmetry evades the no-go

theorem barring unified spacetime and internal symmetries. This theorem (the Coleman–Mandula theorem) can be evaded, since at the time it was written, graded Lie symmetry algebras were not yet considered. For nonzero central charges, the external automorphism algebra becomes a subalgebra of $U(N)$ determined by the requirement that invariant antisymmetric tensors $a^{\ell ij}$ exist.

The representations of the algebra [13]–[14] span an increasing range of spins as the number N of $D=4$ supersymmetries increases. For massive representations without central charges, the spins of the smallest supersymmetry representation extend from states of spin 0 (scalars) up to spin $N/2$; with central charges, the spin range can be shortened down to a minimum range of $N/4$. For massless representations, the range of helicities in a PCT (parity–change–time reversal) symmetric multiplet is from $-N/4$ to $N/4$. This spin range has an important implication for the maximal extension of supersymmetry that can be realized in an interacting supersymmetric field theory, because no interacting theories with a finite set of spins exist for spins >2 . Accordingly, the maximal extension of supersymmetry is $N=8$ for massless theories, and in order to have massive states with spins that do not exceed spin 2 in an $N=8$ theory, the central charges have to be active for maximal multiplet shortening.

The $N=8$ supergravity theory, found by Cremmer and Julia in 1978, is thus the largest possible supergravity in $D=4$ dimensions. It contains the following “spin” range (allowing for a certain imprecision of expression: for massless fields one should really speak only of helicities)

N = 8 supergravity spins

Spin	2	3/2	1	1/2	0
Multiplicity	1	8	28	56	70

In order to realize the automorphism $SU(8)$ symmetry, one has to consider the field strengths for the 28 spin-1 fields, separated into complex self-dual and anti-self-dual parts in their antisymmetric Lorentz indices. These complex field strengths can then be endowed with a complex 28-dimensional representation of $SU(8)$. The 70 scalars, on the other hand, fit precisely into the four-index antisymmetric self-dual representation of $SU(8)$, $\phi^{i_1 i_2 i_3 i_4} = 1/(4!) \epsilon^{i_1 i_2 i_3 i_4 j_1 j_2 j_3 j_4} \bar{\phi}_{j_1 j_2 j_3 j_4}$. It is the use of the eight-index epsilon tensor here that restricts the automorphism group to $SU(8)$ instead of $U(8)$.

The $SU(8)$ automorphism symmetry of $N=8$ supergravity theory is linearly realized. It plays an important role in another symmetry of this theory which is highly nonlinear. This theory has a

remarkable nonlinear E_7 symmetry. In fact, the 70 scalars form a nonlinear sigma model with the fields taking their values in the coset space $E_7/SU(8)$ (of dimension $133 - 63 = 70$), where the $SU(8)$ divisor is the linearly realized automorphism group discussed above.

The extended supergravities point to another aspect of supergravity theory: the existence of higher-dimensional supergravities, from which the extended theories in $D=4$ spacetime can be derived by Kaluza–Klein dimensional reduction. If one considers a D' dimensional massless theory in a spacetime where d dimensions form a compact d -torus, then the theory can be viewed as a $D = D' - d$ dimensional theory in which the discrete Fourier modes arising from the periodicity requirements on the d -torus give rise to towers of equally spaced massive Kaluza–Klein states, plus a massless sector in $D' - d$ dimensions corresponding to the modes with no dependence on the d -torus coordinates.

Importantly, $N=8$ supergravity in four-dimensional spacetime can be obtained in this way from a supergravity theory that exists in 11 spacetime dimensions. Upon dimensional reduction on a 7-torus to four dimensions, one obtains $N=8, D=4$ supergravity at the massless level, plus an infinite tower of massive $N=8$ supermultiplets with central charges so that their spin range extends only up to spin 2. This $D=11$ supergravity was in fact found before the $N=8$ theory by Cremmer, Julia, and Scherk, with the details of the more complicated $N=8, D=4$ theory being worked out via the techniques of Kaluza–Klein dimensional reduction. The fields of the $D=11$ theory include an exotic field type not encountered in $D=4$ theories: the bosonic fields of the theory comprise the graviton e_M^A plus a three-index antisymmetric tensor gauge field C_{MNP} . Counting the number of propagating modes of these fields for a given momentum value gives $44 + 84 = 128$ bosonic degrees of freedom. This precisely balances the 128 fermionic degrees of freedom coming from the $D=11$ gravitino $\psi_{M\alpha}$.

Supergravity Effective Theories, Strings and Branes

The hope for a cancellation of the ultraviolet divergences in a supersymmetric theory of gravity turned out to be ephemeral, although there is in fact a postponement of the divergence onset until a higher order in quantum field loops. There is agreement that the nonmaximal supergravities diverge at the three-loop order. For the $N=8, D=4$ theory, the situation remains unclear,

but divergences are nonetheless expected to occur at some finite loop order.

This persistence of nonrenormalizability in $D=4$ supergravity theories is no longer seen as a disaster, however, because these theories are now seen as effective theories for the massless modes arising from a deeper microscopic quantum theory. In addition, the theories that are most directly connected to this underlying quantum theory are, surprisingly, the maximal supergravities in spacetime dimensions 10 and 11. $D=11$ supergravity can be dimensionally reduced on a 1-torus (i.e., a circle) to $D=10$ where the massless sector yields type IIA supergravity theory. This theory is the effective theory for a consistent quantum theory of type IIA superstrings in $D=10$. Theories of relativistic strings (i.e., one-dimensional extended objects) have strikingly different properties from theories of point particles. In particular, the spread-out nature of the interactions leads to a damping out of the quantum field theory divergences, while the underlying supersymmetry causes a cancellation of other infinities that could have arisen owing to the two-dimensional nature of the string world sheets. This gives, for the first time, a perturbatively well-defined quantum theory including gravity.

In addition to the type IIA theory, there are four other consistent superstring theories in $D=10$, and these are in turn related to various $D=10$ supergravity effective theories for the massless modes: type IIB, $E_8 \times E_8$ heterotic, $SO(32)$ heterotic, and $SO(32)$ type I. Remarkably, the maximal $D=11$ supergravity enters into this picture as well, as a consequence of a pattern of duality symmetries that have been found among the superstring theories.

The dualities of string theory are directly related to the nonlinear symmetries of the dimensionally reduced supergravities in $D=4$. The string quantum corrections do not respect the E_7 symmetry of the classical $N=8$ theory, but they do respect a discrete subgroup of this symmetry in which the E_7 group elements are required to take integer values: $E_7(\mathbb{Z})$.

This quantum-level restriction to a discrete subgroup can be seen from another phenomenon characteristic of superstring theories: the existence of “electric” and “magnetic” brane solutions. The antisymmetric-tensor (or “form”) fields of the higher-dimensional supergravities naturally give rise to solitonic solutions in which $p+1$ dimensions form a flat Poincaré invariant subspace. This can be interpreted as the world volume of an infinite p -brane extended object. In the $D=11$ supergravity theory, the branes that emerge in this way are a 2-brane and a 5-brane. The three-dimensional world volume of the 2-brane naturally couples to the

3-form field C_{MNP} , just as an ordinary Maxwell vector field couples to the one-dimensional world line of a point particle (or 0-brane). The 2-brane is thus naturally electrically charged with respect to the 3-form field; its charge can be obtained, in a direct generalization of the Maxwell case, from a Gauss' law integral of the field strength $H_{[4]} = dC_{[3]}$ over a 7-sphere at spatial infinity in the eight directions transverse to the brane worldvolume. The 5-brane, on the other hand, has a magnetic type charge; it is the 7-form dual to $H_{[4]}$ that is integrated to give its charge. In addition to these static infinite p -branes, the theory contains dynamical finite-extent branes as well, although for these one generally does not have explicit solutions.

As one reduces a higher-dimensional supergravity to lower and lower dimensions, there is a proliferation of solitonic brane solutions of varying dimensionality, and of both electric and magnetic charge types. In a quantum theory context, these electrically and magnetically charged branes pair up in ways that must satisfy a generalization of the Dirac quantization condition for $D=4$ electric and magnetic point particles. This ends up requiring all the supergravity solitonic brane charges to lie on a charge lattice. It is the requirement that this discrete brane-charge lattice be respected that restricts the classical supergravity nonlinear symmetry groups to discrete duality subgroups.

The dualities relate brane solutions within a given theory and also between different string theories. They include transformations that invert the radii of compactifying tori, giving a large–small compactification scale duality. They also include transformations that invert the string coupling constant, thus interchanging strong and weak coupling. The type IIB theory, for example, is self-dual under strong–

weak coupling duality. In the case of the type IIA theory, however, something remarkable happens. The strong coupling limit of this theory turns out to be related by duality, not to another string theory, but to the maximal $D=11$ supergravity. The role of the Kaluza–Klein massive modes for the 11 to 10 reduction is played by an infinite tower of extremal charged black holes.

Thus, even $D=11$ supergravity theory has a role to play in the effective theory of the underlying quantum dynamics. This underlying theory has been dubbed “M-theory.” It is still only partially understood, but many of its most important properties are presaged by the remarkable nonlinear structure of the classical supergravities.

See also: Brane Construction of Gauge Theories; Brane Worlds; Branes and Black Hole Statistical Mechanics; Random Algebraic Geometry, Attractors and Flux Vacua; Renormalization: General Theory; Spinors and Spin Coefficients; Stability of Minkowski Space; Supermanifolds; Superstring Theories; Supersymmetric Particle Models; Symmetries and Conservation Laws; Symmetries in Quantum Field Theory: Algebraic Aspects.

Further Reading

- Buchbinder JL and Kuzenko SM (1998) *Ideas and Methods of Supersymmetry and Supergravity*. Bristol: IoP Publishing Ltd.
 Stelle KS (1998) BPS branes in supergravity, Trieste 1987 School of High-Energy Physics and Cosmology, arXiv:hep-th/9803116.
 Van Nieuwenhuizen P (1981) Supergravity. *Physics Reports* 68: 189–398.
 Wess J and Bagger J (1983) *Supersymmetry and Supergravity*. Princeton: Princeton University Press.

Supermanifolds

F A Rogers, King's College London, London, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

A supermanifold is a generalization of a classical manifold to include coordinates that are in some sense anticommuting. Much of the motivation for the study of supermanifolds comes from supersymmetric physics, where it is useful to have a formalism which treats fermions and bosons in the same way. The underlying reason for the

effectiveness of supermanifolds is that anticommuting coordinates allow the fermionic canonical anticommutation relations to be handled in a way analogous to the bosonic canonical commutation relations. Supersymmetric methods have proved immensely effective in fundamental physics; they also play a considerable role in geometrical index theory in mathematics. In this article we describe supermanifolds from two points of view – geometric and algebraic – and consider some of the standard features of manifold calculus, including integration since this is an area where the distinctive features of this generalized geometry are particularly apparent.

One situation where supermanifolds are used in physics is in the superspace formulation of supergravity, where the physical fields are found in the component fields in the Taylor expansion of functions on the supermanifold in anticommuting variables. More fundamentally, the symmetry groups of supersymmetric theories have commuting and anticommuting generators, and are examples of super Lie groups, which are supermanifolds with a compatible group structure.

Some Algebraic Preliminaries

The coordinates of a supermanifold have particular algebraic features which are best understood by introducing some of the basic concepts of superalgebra. (The word super here does not imply superiority, simply the extension of some classical concept to have odd as well as even, anticommuting as well as commuting, elements.) A “super vector space” is a vector space V together with a direct sum decomposition

$$V = V_0 \oplus V_1 \quad [1]$$

The subspaces V_0 and V_1 are referred to, respectively, as the even and odd parts of V . A general element v of V thus has the unique decomposition $v = v_0 + v_1$ with v_0 in V_0 and v_1 in V_1 . We will normally consider homogeneous elements, that is, elements v which are either even or odd, with parity denoted by $|v|$, so that $|v| = i$ if v is in $V_i, i = 0, 1$. (Arithmetic of parity indices $i = 0, 1$ is always modulo 2.) A superalgebra is a super vector space whose elements can be multiplied together in such a way that the product of an even element with an even element and that of an odd element with an odd element are both even, while the product of an odd element with an even element is odd; more formally:

Definition 1

- (i) A “superalgebra” is a super vector space $A = A_0 \oplus A_1$ which is also an algebra which satisfies $A_i A_j \subset A_{i+j}$.
- (ii) The superalgebra is “supercommutative” if, for all homogeneous a, b in $A, ab = (-1)^{(|a||b|)} ba$.

If the algebra is supercommutative then odd elements anticommute, and the square of an odd element is zero. The basic supercommutative superalgebra used is the real Grassmann algebra with generators $1, \beta_1, \beta_2, \dots$ and relations

$$1\beta_i = \beta_i 1 = \beta_i, \quad \beta_i \beta_j = -\beta_j \beta_i \quad [2]$$

A typical element of this algebra is then

$$a = a_0 1 + \sum_i a_i \beta_i + \sum_{i < j} a_{ij} \beta_i \beta_j \cdots \quad [3]$$

This algebra, which is denoted R_S , is a superalgebra with $R_S := R_{S,0} \oplus R_{S,1}$, where $R_{S,0}$ consists of linear combinations of products of even numbers of the anticommuting generators, while $R_{S,1}$ is built similarly from odd products.

The Grassmann algebra R_S is used to build the (m, n) -dimensional superspace $R_S^{m,n}$ in the following way:

Definition 2. An (m, n) -dimensional superspace is the space

$$R_S^{m,n} = \underbrace{R_{S,0} \times \cdots \times R_{S,0}}_{m \text{ copies}} \times \underbrace{R_{S,1} \times \cdots \times R_{S,1}}_{n \text{ copies}} \quad [4]$$

A typical element of $R_S^{m,n}$ is written as $(x^1, \dots, x^m; \xi^1, \dots, \xi^n)$, where the convention is used that lower case Latin letters represent even objects and lower case Greek letters represent odd objects, while small capitals are used for objects of mixed or unspecified parity.

As will be described in more detail below, in the geometric approach supermanifolds are spaces locally modeled on $R_S^{m,n}$. In order to define a supermanifold, we will need to define a topology on this space, and to have some notion of differentiation. Consider first multilinear functions of purely anticommuting variables. If there are n such variables, ξ^1, \dots, ξ^n , then a multilinear function F can be expressed in the form

$$F(\xi^1, \dots, \xi^n) = F_0 + \sum_{i=1}^n F_i \xi^i + \sum_{1 \leq i < j}^n F_{ij} \xi^i \xi^j + \cdots + F_{1\dots n} \xi^1 \cdots \xi^n \quad [5]$$

where the coefficients F_0, F_i and so on are real numbers. Such functions will be known (anticipating the terminology for functions of both odd and even variables) as supersmooth. (A useful notation will be to write

$$F(\xi^1, \dots, \xi^n) = \sum_{\mu} F_{\mu} \xi^{\mu} \quad [6]$$

with μ a multi-index $\mu = \mu_1 \cdots \mu_k$ and $\xi^{\mu} = \xi^{\mu_1} \cdots \xi^{\mu_k}$. The set of multi-indices is restricted to those where $1 \leq \mu_1 < \cdots < \mu_k \leq n$.) More general supersmooth functions, with the coefficients F_0, \dots taking values in \mathbb{C}, R_S , or some other algebra are also possible.

Differentiation of supersmooth functions of anti-commuting variables is defined by linearity together with the rule

$$\frac{\partial(\xi^{\mu_1}\xi^{\mu_2}\dots\xi^{\mu_r})}{\partial\xi^j} = \begin{cases} (-1)^{k-1}\xi^{\mu_1}\dots\widehat{\xi^{\mu_k}}\dots\xi^{\mu_r} & \text{if } j = \mu_k \\ 0 & \text{otherwise} \end{cases} \quad [7]$$

where the caret $\widehat{}$ indicates an omitted factor.

In order to extend the notion of supersmoothness to functions on the more general superspace $\mathbb{R}_S^{m,n}$, we should strictly take note of the fact that an even Grassmann variable is not simply a real or complex variable, as explained in the appendix. Assuming this done, a supersmooth function on the general superspace $\mathbb{R}_S^{m,n}$ can then be defined as a function of the form

$$F(x^1, \dots, x^m, \xi^1, \dots, \xi^n) = \sum_{\mu} F_{\mu}(x^1, \dots, x^m) \xi^{\mu} \quad [8]$$

with each coefficient function F_{μ} a smooth function on \mathbb{R}^m .

The final preparatory idea needed is the topology on the superspace $\mathbb{R}_S^{m,n}$. It turns out that a coarse, non-Hausdorff topology leads to most of the supermanifolds used in physics. In order to define this topology, we introduce a mapping

$$\epsilon : \mathbb{R}_S \rightarrow \mathbb{R}$$

defined by

$$\epsilon\left(a_0\mathbf{1} + \sum_i a_i\beta_i + \sum_{i<j} a_{ij}\beta_i\beta_j \dots\right) = a_0 \quad [9]$$

and the related mapping

$$\epsilon : \mathbb{R}_S^{m,n} \rightarrow \mathbb{R}^m$$

defined by

$$\epsilon((x^1, \dots, x^m; \xi^1, \dots, \xi^n)) = (\epsilon(x^1), \dots, \epsilon(x^m)) \quad [10]$$

These maps project out all the nilpotent Grassmann generators, leaving simply the real part. The topology involves the inverse of these projection maps: a subset U of $\mathbb{R}_S^{m,n}$ is said to be open if and only if there exists an open set V in \mathbb{R}^m such that $U = \epsilon^{-1}(V)$. Thus, an open set is unlimited in the nilpotent directions.

In the sequel, where we consider integration, the superdeterminant of the matrix M of an endomorphism of a super vector space V will be useful. If V is an (m, n) -dimensional super vector space

(so that V_0 has dimension m and V_1 dimension n), then M will have the block diagonal form

$$\begin{pmatrix} M_{00} & M_{01} \\ M_{10} & M_{11} \end{pmatrix}$$

where the entries of M_{00} and M_{11} are even, whereas those of M_{10} and M_{01} are odd. If $N = M^{-1}$ has block form

$$\begin{pmatrix} N_{00} & N_{01} \\ N_{10} & N_{11} \end{pmatrix}$$

then the superdeterminant of M is defined by

$$\text{S det } M = \det M_{00} \det N_{11}$$

It can be shown that the superdeterminant obeys the product rule, unlike the obvious generalization of the determinant to the super case.

The Geometric Approach to Supermanifolds

A manifold is a space locally modeled on the topological space \mathbb{R}^m , where m is the dimension of the manifold. Thus, each point in a manifold has a neighborhood which is essentially a neighborhood in \mathbb{R}^m . The most geometrically intuitive approach to supermanifolds is to generalize this directly by modeling a space locally on an extension of \mathbb{R}^m to include anticommuting variables; the most straightforward space with the required algebraic property is the superspace $\mathbb{R}_S^{m,n}$ built from a Grassmann algebra, leading to a supermanifold of dimension (m, n) . (The dimension of a supermanifold is a pair of integers, indicating the numbers of even and odd coordinates of each point.)

The formal definition of a supermanifold will now be given in a manner very closely analogous to that of a classical manifold.

Definition 3. Let M be a set.

- (i) An (m, n) open chart on M is a pair (U, ϕ) such that U is a subset of M and ϕ is an injective map of U into $\mathbb{R}_S^{m,n}$, with the image $\phi(U)$ an open set in $\mathbb{R}_S^{m,n}$.
- (ii) An (m, n) atlas on M is a collection $\{(U_{\alpha}, \phi_{\alpha})\}$ of (m, n) charts on M such that the U_{α} cover M and, whenever $U_{\alpha} \cap U_{\beta}$ is not empty, the change of coordinate function $\phi_{\alpha} \circ \phi_{\beta}^{-1}$ is supersmooth.

An (m, n) -dimensional supermanifold is a set M together with a maximal (m, n) atlas on M .

The space M is given a topology by defining $U \subset M$ to be open if and only if, for each α such that $U \cap U_{\alpha}$ is not empty, the set $\phi_{\alpha}(U \cap U_{\alpha})$ is an open subset of $\mathbb{R}_S^{m,n}$.

Examples of supermanifolds include $\mathbb{R}_S^{m,n}$ itself, and also supermanifolds constructed from the data of a vector bundle over a classical manifold in a manner which will now be described. If N is a classical m -dimensional real manifold and E is an n -dimensional vector bundle over N , then an (m, n) -dimensional supermanifold can be constructed in the following way: suppose that $\{(V_\alpha, \psi_\alpha)\}$ is an atlas of charts on N , so that each V_α is an open subset of N and each ψ_α is an injective map of V_α onto an open subset of \mathbb{R}^m , with $\psi_\alpha \circ \psi_\beta^{-1}$ smooth. Suppose further that the V_α are also local trivialization neighborhoods of the bundle E with transition functions $g_{\alpha\beta}: V_\alpha \cap V_\beta \rightarrow GL(n)$. Then we build the supermanifold M by patching together the sets $\epsilon^{-1}(\psi_\alpha(V_\alpha) \times \mathbb{R}_S^{0,n})$ in a consistent way. This leads to a supermanifold with coordinate change functions

$$\begin{aligned} \phi_\alpha \circ \phi_\beta^{-1}(x_\beta^1, \dots, x_\beta^m, \xi_\beta^1, \dots, \xi_\beta^n) \\ = (x_\alpha^1, \dots, x_\alpha^m, \xi_\alpha^1, \dots, \xi_\alpha^n) \end{aligned}$$

where

$$\begin{aligned} (x_\alpha^1, \dots, x_\alpha^m) &= \psi_\alpha \circ \psi_\beta^{-1}(x_\beta^1, \dots, x_\beta^m) \\ \xi_\alpha^j &= \sum_{k=1}^n g_{\alpha\beta}^{jk}(x_\beta^1, \dots, x_\beta^m) \xi_\beta^k \end{aligned} \quad [11]$$

(Here again we refer to the appendix for the way in which functions of even Grassmann variables, as opposed simply to real numbers, are handled.) Particular examples of this construction are the tangent bundle over N and bundles of spinors over N . It was actually shown by Batchelor that all real, supersmooth supermanifolds are of this form.

A similar definition may be made of a complex supermanifold using a complex Grassmann algebra, with the coordinate transition functions required to be superanalytic. In this case, supermanifolds which are not related to vector bundles in the manner described above are possible, basically because partitions of unity do not exist in the analytic setting. An example is the twisted supertorus, which is built over the standard torus and has transition functions $(z, \zeta) \rightarrow (z+1, \zeta)$ and $(z, \zeta) \rightarrow (z+a+\alpha\zeta, \zeta+\alpha)$, extending the standard torus with transition functions $z \rightarrow z+1, z \rightarrow z+a$. (Here a, α are, respectively, even and odd constants.) This supermanifold is an example of a super Riemann surface; such surfaces play an important role in the quantization of the spinning string.

As with classical manifolds, a natural class of functions can be defined on a supermanifold: a function f on an open subset U of the

supermanifold M is said to be supersmooth if, for each α such that $U \cap U_\alpha$ is nonempty, the function $f \circ \phi_\alpha^{-1}$ is supersmooth on $\phi_\alpha(U \cap U_\alpha)$. In local coordinates supersmooth functions are such that $f(x^1, \dots, x^m, \xi^1, \dots, \xi^n) = \sum_\mu f_{\alpha\mu}(x^1, \dots, x^m) \xi^\mu$ with each $f_{\alpha\mu}$ a smooth function.

The Algebraic Approach to Supermanifolds

In the algebraic approach to supermanifolds, it is the algebra of functions, rather than the manifold itself, which is extended to include anticommuting elements. In this approach an (m, n) -dimensional supermanifold is defined to be a pair (N, A) , where N is an m -dimensional classical manifold and A is a sheaf of superalgebras over N with various properties, described below. The statement that A is a sheaf of algebras over N means that corresponding to each open subset U of N there is an algebra $A(U)$; also, if $V \subset U$, there is a “restriction map” $\rho_{U,V}$ mapping $A(U)$ into $A(V)$, and the various restriction maps obey certain consistency conditions. A particular example of such a sheaf (with trivial odd part) is the sheaf A_0 of real-valued functions on N , with $A_0(U) = C^\infty(U)$, the set of real-valued smooth functions on U and $\rho_{U,V}$ mapping a function in $C^\infty(U)$ to its restriction in $C^\infty(V)$. The defining property of the sheaf corresponding to an (m, n) -dimensional supermanifold is that there is a cover $\{U_\alpha\}$ of N for which the algebras $A(U_\alpha)$ have the form $A(U_\alpha) \cong C^\infty(U_\alpha) \otimes \Lambda(\mathbb{R}^n)$, so that a typical element f of $A(U_\alpha)$ may be expressed as $f = \sum_\mu f_\mu \xi^\mu$, where $f_\mu \in C^\infty(U_\alpha)$ and ξ^1, \dots, ξ^n are generators of $\Lambda(\mathbb{R}^n)$. The notation here is chosen to emphasize the close correspondence with the algebra of smooth functions described at the end of the previous section. This makes it clear that, despite an apparent difference, the two approaches lead to essentially equivalent supermanifolds.

The advantage of the algebraic approach is its mathematical elegance and economy – there is no need to introduce the auxiliary Grassmann algebra \mathbb{R}_S in which coordinate functions take values – but from the point of view of physicists, the geometric point of view has two advantages: first, it is closer to the standard manifold picture and thus easier to grasp, and, second, it allows a wider class of supermanifolds, because Grassmann constants are allowed; for instance, the twisted supertorus described above cannot be included in the algebraic approach without either introducing an auxiliary algebra or moving to the more difficult concept of a family of supermanifolds.

While there have been various attempts to develop infinite-dimensional supermanifolds, most of the constructions have been developed for very specific purposes, such as path integration and functional integration methods for theories with fermions. Even the question of defining a basic infinite-dimensional superalgebra with the necessary analytic properties, such as a Hilbert–Banach superalgebra, requires sophisticated procedures, so that the development of a theory of infinite-dimensional supermanifolds becomes extremely technical.

Calculus on Supermanifolds

Much of the calculus of functions on supermanifolds proceeds in simple analogy to that of classical manifolds, with addition sign factors occurring whenever two odd quantities are transposed. For instance, a vector field on M may be described as a superderivation of the algebra of supersmooth functions on M , that is, a linear mapping of this space obeying the super Leibnitz rule $Xfg = Xf g + (-1)^{(|X||f|)} f Xg$. Standard examples of vector fields (defined locally) are coordinate derivatives $\partial/\partial x^i$ and $\partial/\partial \xi^j$, defined by $(\partial/\partial x^i)f = \partial_i(f \circ \phi)$ and $(\partial/\partial \xi^j)f = \partial_{j+m}(f \circ \phi)$ with ϕ the coordinate function corresponding to the coordinates $(x^1, \dots, x^m; \xi^1, \dots, \xi^n)$. Equipped with this concept of vector field, much of differential calculus on manifolds can be directly generalized to supermanifolds in a relatively straightforward way. However, in the case of integration the situation is quite different. The standard approach to integration of anticommuting variables is the Berezin integral, which is a formal, algebraic integral that is not an antiderivative and has no measure-theoretic features. There are various reasons why such an integral is used: for instance, even the simple function ξ of a single anticommuting variable has no antiderivative, while the topology on $\mathbb{R}_S^{m,n}$ does not allow open sets which discriminate in odd directions. Additionally, when changing variables on $\mathbb{R}_S^{m,n}$ it is the superdeterminant of the Jacobian matrix which must be used. In the purely odd sector, differentials thus transform the “wrong” way.

The Berezin integral of a function f of n anticommuting variables is defined by

$$\int d^n \xi \left(\sum_{\mu} f_{\mu} \xi^{\mu} \right) = f_{1\dots n} \quad [12]$$

In other words, Berezin integration simply picks out the coefficient of the highest-order term, thus resembling differentiation more than integration in the classical sense. Nonetheless, the Berezin integral has very useful properties, in particular allowing direct analogues of Fourier transformations and

integral kernel. Given that it is the algebra of functions, and the operators acting on these algebras, which is the key element in supergeometry, these are vital properties of the integral.

The transformation rule under change of variable is the inverse of that which one expects. For instance, in the case of a single variable, if one makes the transformation $\xi \rightarrow \phi = a\xi + \beta$ with a and β constants, a direct calculation shows that the integral is invariant provided that one sets $d\xi = a d\phi$.

Integration on $\mathbb{R}_S^{m,n}$ is essentially defined by combining classical integration for the even variables with Berezin integration for odd variables, giving

$$\begin{aligned} \int_{\epsilon^{-1}(V)} d^m x d^n \xi \left(\sum_{\mu} f_{\mu}(x^1, \dots, x^m) \xi^{\mu} \right) \\ = \int_V d^m x (f_{1\dots n}(x^1, \dots, x^m)) \end{aligned} \quad [13]$$

This also defines integration on supermanifolds, provided that we can find a rule for the change of variable. This, as indicated above, may be done by using the superdeterminant of the Jacobian matrix. Suppose that (y, ϕ) are a new set of coordinates on our supermanifold. Then an invariant definition of integral is obtained if we set

$$d^m y d^n \xi = \text{Sdet} \begin{pmatrix} \frac{\partial y}{\partial x} & \frac{\partial y}{\partial \xi} \\ \frac{\partial \phi}{\partial x} & \frac{\partial \phi}{\partial \xi} \end{pmatrix} d^m x d^n \xi \quad [14]$$

Appendix

We now describe the device which allows functions of even Grassmann variables to be handled simply as functions of conventional variables. The necessary class of functions is captured by defining supersmooth functions on $\mathbb{R}_S^{m,0}$ as extensions by Taylor expansion from smooth functions on \mathbb{R}^m .

Definition 4. The function $F: \mathbb{R}_S^{m,0} \rightarrow \mathbb{R}_S$ is said to be supersmooth if there exists a smooth function $\tilde{F}: \mathbb{R}^m \rightarrow \mathbb{R}$, such that

$$\begin{aligned} F(x^1, \dots, x^m) \\ = \tilde{F}(\epsilon(x)) + \sum_{i=1}^m (x^i - \epsilon(x^i)1) \frac{\partial \tilde{F}}{\partial x^i}(\epsilon(x)) \\ + \frac{1}{2} \sum_{i,j=1}^m (x^i - \epsilon(x^i)1) \\ \times (x^j - \epsilon(x^j)1) \frac{\partial^2 \tilde{F}}{\partial x^i \partial x^j}(\epsilon(x)) \dots \end{aligned} \quad [15]$$

(Although this Taylor series will in general be infinite, it gives well-defined coefficients for each

β_μ in the expansion [3], so that the value of F is a well-defined element of \mathbb{R}_S .) A number of different classes of function can be obtained, by varying the space in which the function \tilde{F} takes its value.

See also: Batalin–Vilkovisky Quantization; BRST Quantization; Graded Poisson Algebras; Path-Integrals in Non Commutative Geometry; Random Matrix Theory in Physics; Supergravity; Superstring Theories; Supersymmetric Particle Models; Supersymmetric Quantum Mechanics.

Further Reading

- Batchelor M (1979) The structure of supermanifolds. *Transactions of the American Mathematical Society* 253: 329–338.
- Batchelor M (1980) Two approaches to supermanifolds. *Transactions of the American Mathematical Society* 258: 257–270.
- Berezin FA (1987) *Introduction to Superanalysis*. Dordrecht: Reidel.
- Berezin FA and Leites DA (1976) Supermanifolds. *Soviet Maths Doklady* 16: 1218–1222.
- Crane L and Rabin JM (1988) Super Riemann surfaces: uniformization and Teichmüller theory. *Communications in Mathematical Physics* 113: 601–623.

- DeWitt BS (1992) *Supermanifolds*. Cambridge: Cambridge University Press.
- Howe PS (1979) Super Weyl transformations in two dimensions. *Journal of Physics A* 12: 393–402.
- Jadczyk A and Pilch K (1981) Superspace and supersymmetries. *Communications in Mathematical Physics* 78: 373–390.
- Kostant B (1977) Graded manifolds, graded Lie theory and prequantization. In: *Differential Geometric Methods in Mathematical Physics*, Lecture Notes in Mathematics, Springer.
- Kupsch J and Smolyanov O (2000) Hilbert norms for graded algebras. *Proceedings of the American Mathematical Society* 128: 1647–1653.
- Polchinski J (1998) *String Theory*, vol. II. Cambridge: Cambridge University Press.
- Rogers A (2003) Supersymmetry and Brownian motion on supermanifolds. *Infinite Dimensional Analysis, Quantum Probability and Related Topics* 6(suppl. 1): 83–102.
- Salam A and Strathdee J (1974) Super-gauge transformations. *Nuclear Physics B* 76: 477–482.
- Voronov AA (1992) Geometric integration theory on supermanifolds. *Soviet Scientific Reviews C: Mathematical Physics Reviews* 9: 1–138.
- Wess J and Bagger J (1983) *Supersymmetry and Supergravity*. Princeton: Princeton University Press.
- West P (1990) *Introduction to Supersymmetry and Supergravity*. Singapore: World Scientific.

Superstring Theories

C Bachas and J Troost, Ecole Normale Supérieure, Paris, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction

String theory postulates that all elementary particles in nature correspond to different vibration states of an underlying relativistic string. In the quantum theory both the frequencies and the amplitudes of vibration are quantized, so that the quantum states of a string are discrete. They can be characterized by their mass, spin, and various gauge charges. One of these states has zero mass and spin equal to $2\hbar$, and can be identified with the messenger of gravitational interactions, the graviton. Thus, string theory is a candidate for a unified theory of all fundamental interactions, including quantum gravity.

In this article, we discuss the theory of superstrings as consistent theories of quantum gravity. The aim is to provide a quick (mostly lexicographic and bibliographic) entry to some of the salient features of the subject for a nonspecialist audience. Our treatment is thus neither complete nor comprehensive – there exist for this several excellent expert books, in particular

by Green, *et al.* (1987) and by Polchinski (1998). An introductory textbook by Zwiebach (2004) is also highly recommended for beginners. Several other complementary reviews on various aspects of superstring theories are available on the internet (see the “Further reading” section); some more will be given as we proceed.

The Five Superstring Theories

Theories of relativistic extended objects are tightly constrained by anomalies, that is, quantum violations of classical symmetries. These arise because the classical trajectory of an extended p -dimensional object (or “ p -brane”) is described by the embedding $X^\mu(\zeta^a)$, where $\zeta^a=0,\dots,p$ parametrize the brane world volume, and $X^\mu=0,\dots,D-1$ are coordinates of the target space. The quantum mechanics of a single p -brane is therefore a $(p+1)$ -dimensional quantum field theory, and as such suffers *a priori* from ultraviolet divergences and anomalies. The case $p=1$ is special in that these problems can be exactly handled. The story for higher values of p is much more complicated, as will become apparent later on.

The theory of ordinary loops in space is called closed bosonic string theory. The classical trajectory

of a bosonic string extremizes the Nambu–Goto action (proportional to the invariant area of the world sheet)

$$S_{\text{NG}} = -\frac{1}{2\pi\alpha'} \int d^2\zeta \sqrt{-\det(G_{\mu\nu}\partial_a X^\mu \partial_b X^\nu)} \quad [1]$$

where $G_{\mu\nu}(X)$ is the target-space metric, and α' is the Regge slope (which is inversely proportional to the string tension and has dimensions of length squared). In flat spacetime, and for a conformal choice of world-sheet parameters $\zeta^\pm = \zeta^0 \pm \zeta^1$, the equations of motion read:

$$\partial_+ \partial_- X^\mu = 0 \quad \text{and} \quad \eta_{\mu\nu} \partial_\pm X^\mu \partial_\pm X^\nu = 0 \quad [2]$$

with $\eta_{\mu\nu}$ the Minkowski metric. The X^μ are thus free two-dimensional fields, subject to quadratic phase-space constraints known as the Virasoro conditions. These can be solved consistently at the quantum level in the critical dimension $D=26$. Otherwise, the symmetries of eqns [2] are anomalous: either Lorentz invariance is broken, or there is a conformal anomaly leading to unitarity problems. (For $D < 26$, unitary noncritical string theories in highly curved rather than in the originally flat background can be constructed.)

Even for $D=26$, bosonic string theory is, however, sick because its lowest-lying state is a tachyon, that is, it has negative mass squared. This follows from the zeroth-order Virasoro constraints,

$$m^2 = -p^M p_M = \frac{4}{\alpha'} (N_L - 1) = \frac{4}{\alpha'} (N_R - 1) \quad [3]$$

where N_L (N_R) is the sum of the frequencies of all left(right)-moving excitations on the string world sheet. The negative contribution to m^2 comes from quantum fluctuations, and is analogous to the well-known Casimir energy. The tachyon has $N_L = N_R = 0$. Its presence signals an instability of Minkowski spacetime, which in bosonic string theory is expected to decay, possibly to some lower-dimensional highly curved geometry. The details of how this happens are not, at present, well understood.

The problem of the tachyon is circumvented by endowing the string with additional, anticommuting coordinates, and requiring spacetime supersymmetry. This is a symmetry that relates string states with integer spin, obeying Bose–Einstein statistics, to states with half-integer spin obeying Fermi–Dirac statistics. There exist two standard descriptions of the superstring: the Ramond–Neveu–Schwarz (RNS) formulation, where the anticommuting coordinates ψ^μ carry a spacetime vector index, and the Green–Schwarz (GS) formulation in which they transform as a spacetime spinor θ^α . Each has its advantages and

drawbacks: the RNS formulation is simpler from the world sheet point of view, but awkward for describing spacetime fermionic states; in the GS formulation, on the other hand, spacetime supersymmetry is manifest but quantization can only be carried out in the restrictive light-cone gauge. A third formulation, possibly combining the advantages of the other two, has been proposed more recently by Berkovits (2002) – it is still being developed.

Anomaly cancelation leads to five consistent superstring theories, all defined in $D=10$ flat spacetime dimensions. They are referred to as type IIA, type IIB, heterotic $\text{SO}(32)$, heterotic $E_8 \times E_8$, and type I. The two type II theories are given (in the RNS formulation) by a straightforward extension of eqns [2]:

$$\partial_+ \partial_- X^\mu = \partial_\mp \psi_\pm^\mu = 0 \quad \text{and} \quad \eta_{\mu\nu} \psi_\pm^\mu \partial_\pm X^\nu = 0 \quad [4]$$

The left- and right-moving world sheet fermions can be separately periodic or antiperiodic – these are known as Ramond (R) and Neveu–Schwarz (NS) boundary conditions. Ramond fermions have zero modes obeying a Dirac γ -matrix algebra, and which must thus be represented on spinor space. As a result, out of the four possible boundary conditions for ψ_+^μ and ψ_-^μ , namely NS–NS, R–R, NS–R, or R–NS, the first two give rise to string states that are spacetime bosons, while the other two give rise to states that are spacetime fermions. Consistency of the theory further requires that one only keep states of definite world-sheet fermion parities – an operation known as the Gliozzi–Scherk–Olive (GSO) projection. This operation removes the would-be tachyon, and acts as a chirality projection on the spinors. The type IIA and IIB theories differ only in that the spinors coming from the left and right Ramond sectors have the opposite chirality in type IIA and the same chirality in type IIB.

The fact that string excitations split naturally into noninteracting left and right movers is crucial for the construction of the heterotic strings. The key idea is to put together the left-moving sector of the $D=10$ type II superstring and the right-moving sector of the $D=26$ bosonic string. A subtlety arises because the left–right asymmetry may lead to extra anomalies, under global reparametrizations of the string world sheet. These are known as modular anomalies, and we will come back to them in the following section. Their cancelation imposes stringent constraints on the zero modes of the unmatched (chiral) bosons in the right-moving sector. The free-field expansion of these bosons can be written as:

$$X(\zeta^-) = x_R + \alpha' p_R \zeta^- + \sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{i}{n} a_n e^{-2in\zeta^-} \quad [5]$$

where bold-face letters denote 16-component vectors. Modular invariance then requires that the generalized momentum p_R take its values in a sixteen-dimensional, even self-dual lattice. There exist two such lattices, and they are generated by the roots of the Lie groups $\text{Spin}(32)/Z_2$ and $E_8 \times E_8$. They give rise to the two consistent heterotic string theories.

In contrast to the type II and heterotic theories, which are based on oriented closed strings, the type I theory has unoriented closed strings as well as open strings in its perturbative spectrum. The closed strings are the same as in type IIB, except that one only keeps those states that are invariant under orientation reversal ($\zeta^+ \leftrightarrow \zeta^-$). Open strings must also be invariant under this flip, and can furthermore carry pointlike (Chan–Paton) charges at their two endpoints. This is analogous to the flavor carried by quarks at the endpoints of the chromo-electric flux tubes in QCD. Ultraviolet finiteness requires that the Chan–Paton charges span a 32-dimensional vector space, so that open strings transform in bifundamental symmetric or antisymmetric representations of $\text{SO}(32)$. For a thorough review of type I string theory, see the reference Angelantonj and Sagnotti (2002, 2003).

Interactions and Effective Theories

Strings interact by splitting or by joining at a point, as is illustrated in Figure 1. This is a local interaction that respects the causality of the theory. To compute scattering amplitudes, one sums over all world sheets with a given set of asymptotic states, and weighs each local interaction with a factor of the string coupling constant λ . The expansion in powers of λ is analogous to the Feynman-diagram expansion of point-particle field theories. These latter are usually defined by a Lagrangian, or more exactly by a functional-integral measure, and they make sense both for off-shell quantities as well as at the nonperturbative level. In contrast, our current formulation of superstring theory is in terms of a perturbatively defined S -matrix. The advent of dualities has offered glimpses of an underlying nonperturbative structure called M-theory, but

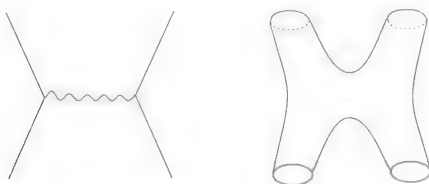


Figure 1 A four-particle and a four-string interaction.

defining it precisely is one of the major outstanding problems in the subject. (One approach consists in trying to define a second-quantized string field theory; see String Field Theory).

Another important expansion of string theory, very useful when it comes to extracting spacetime properties, is in terms of the characteristic string length $l_s = \sqrt{\alpha'}$. At energy scales $El_s \ll 1$, only a handful of massless string states propagate, and their interactions are governed by an effective low-energy Lagrangian. In the type II theories, the massless bosonic states (or rather their corresponding fields) consist of the metric $G_{\mu\nu}$, a scalar field Φ called the dilaton, and a collection of antisymmetric n -form fields coming from both the NS–NS and the R–R sectors. For type IIA, these latter are an NS–NS 2-form B_2 , an R–R 1-form C_1 , and an R–R 3-form C_3 . The leading-order action for these fields reads:

$$S_{\text{IIA}} = \frac{1}{2\kappa^2} \int d^{10}x \left[\sqrt{-G} e^{-2\Phi} (R + 4\partial_\mu \Phi \partial^\mu \Phi - \frac{1}{2}|H_3|^2) - \sqrt{-G} (\frac{1}{2}|F_2|^2 + \frac{1}{2}|F_4 - C_1 \wedge H_3|^2) - \frac{1}{2}B_2 \wedge F_4 \wedge F_4 \right] \quad [6]$$

where $F_2 = dC_1$, $H_3 = dB_2$, and $F_4 = dC_3$ are field strengths, the wedge denotes the exterior product of forms, and $|F_n|^2 = (1/n!) F_{\mu_1 \dots \mu_n} F^{\mu_1 \dots \mu_n}$. The dimensional coupling κ can be expressed in terms of the string-theory parameters, $2\kappa^2 = (2\pi)^7 \lambda^2 \alpha'^4$. A similar expression can be written for the IIB theory, whose R–R sector contains a 0-form, a 2-form, and a 4-form potential, the latter with self-dual field strength.

The action [6], together with its fermionic part, defines the maximally supersymmetric nonchiral extension of Einstein’s gravity in ten dimensions called type IIA supergravity (see Supergravity and Salam and Sezgin (1989)). The dilaton and all antisymmetric tensor fields belong to the supermultiplet of the graviton – they provide together the same number of (bosonic) states as a ten-dimensional nonchiral gravitino. Supersymmetry fixes furthermore completely all two-derivative terms of the action, so that the theory defined by [6] is (almost) unique. (There exists in fact a massive extension of IIA supergravity, which is the low-energy limit of string theory with a nonvanishing R–R 10-form field strength.) It is, therefore, not surprising that it should emerge as the low-energy limit of the (nonchiral) superstring theory. The latter provides, however, an ultraviolet completion of an otherwise nonrenormalizable theory, a completion which is, at least perturbatively, finite and consistent.

The finiteness of string perturbation theory has been, strictly speaking, only established up to two loops – for a recent review see D'Hoker and Phong (2002). However, even though the technical problem is open and hard, the qualitative case for all-order finiteness is convincing. It can be illustrated with the torus diagram which makes a one-loop contribution to string amplitudes. The thin torus of **Figure 2** could be traced either by a short, light string propagating (virtually) for a long time, or by a long, heavy string propagating for a short period of time. In conventional field theory, these two virtual trajectories would have made distinct contributions to the amplitude, one in the infrared and the second in the ultraviolet region. In string theory, on the other hand, they are related by a modular transformation (that exchanges ζ^0 with ζ^1) and must not, therefore, be counted twice. A similar kind of argument shows that all potential divergences of string theory are infrared – they are therefore kinematical (i.e., occur for special values of the external momenta), or else they signal an instability of the vacuum and should cancel if one expands around a stable ground state.

The low-energy limit of the heterotic and type I string theories is $N=1$ supergravity plus super Yang–Mills. In addition to the $N=1$ graviton multiplet, the massless spectrum now also includes gauge bosons and their associated gauginos. The two-derivative effective action in the heterotic case reads:

$$S_{\text{het}} = \frac{1}{2\kappa^2} \int d^{10}x \sqrt{-G} e^{-2\Phi} \times \left[R + 4\partial_\mu \Phi \partial^\mu \Phi + \frac{\kappa^2}{g_{\text{YM}}^2} \text{tr}(F_{\mu\nu} F^{\mu\nu}) - \frac{1}{2} \left| dB_2 - \frac{\kappa^2}{g_{\text{YM}}^2} \omega_3^{\text{gauge}} \right|^2 \right] + \text{fermions} \quad [7]$$

where $\omega_3^{\text{gauge}} = \text{tr}(AdA + (2/3)A^3)$ is the Chern–Simons gauge 3-form. Again, supersymmetry fixes completely the above action – the only freedom is in the choice of the gauge group and of the Yang–Mills

coupling g_{YM} . Thus, up to redefinitions of the fields, the type I theory has necessarily the same low-energy limit.

The $D=10$ supergravity plus super Yang–Mills has a hexagon diagram that gives rise to gauge and gravitational anomalies, similar to the triangle anomaly in $D=4$. It turns out that for the two special groups $E_8 \times E_8$ and $SO(32)$, the structure of these anomalies is such that they can be canceled by a combination of local counter-terms. One of them is of the form $\int B_2 \wedge X_8(F, R)$, where X_8 is an 8-form quartic in the curvature and/or Yang–Mills field strength. The other is already present in the lower line of expression [7], with the replacement $\omega_3^{\text{gauge}} \rightarrow \omega_3^{\text{gauge}} - \omega_3^{\text{Lorentz}}$, where the second Chern–Simons form is built out of the spin connection. Note that these modifications of the effective action involve terms with more than two derivatives, and are not required by supersymmetry at the classical level. The discovery by Green and Schwarz that string theory produces precisely these terms (from integrating out the massive string modes) was called the “first superstring revolution.”

D-Branes

A large window into the nonperturbative structure of string theory has been opened by the discovery of D(irichlet)-branes, and of strong/weak-coupling duality symmetries. A Dp brane is a solitonic p -dimensional excitation, defined indirectly by the property that open string endpoints can attach to its world volume (see **Figure 3**). Stable Dp branes exist in the type IIA and type IIB theories for p even, respectively, odd, and in the type I theory for $p=1$ and 5. They are charged under the R–R $(p+1)$ -form potential or, for $p > 4$, under its magnetic dual. Strictly speaking, only for $0 \leq p \leq 6$ do D-branes resemble regular solitons the word stands for “solitary waves”). The D7 branes are more like

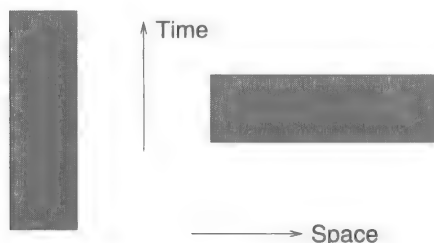


Figure 2 The same torus diagram viewed in two different channels.

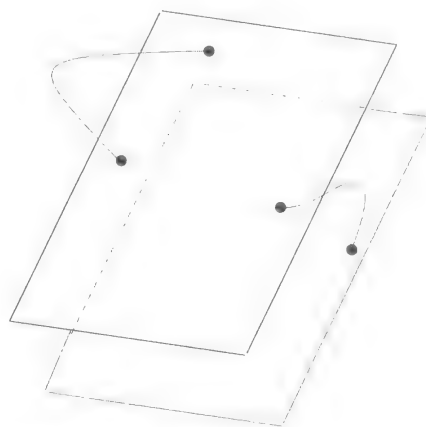


Figure 3 D-branes and open strings.

cosmic strings, the D8 branes are domain walls, while the D9 branes are spacetime filling. Indeed, type I string theory can be thought as arising from type IIB through the introduction of an orientifold 9-plane (required for tadpole cancelation) and of 32 D9 branes.

The low-energy dynamics of a D p brane is described by a supersymmetric abelian gauge theory, reduced from ten down to $p + 1$ dimensions. The gauge field multiplet includes $9 - p$ real scalars, plus gauginos in the spinor representation of the R-symmetry group $SO(9 - p)$. These are precisely the massless states of an open string with endpoints moving freely on a hyperplane. The real scalar fields are Goldstone modes of the broken translation invariance, that is, they are the transverse coordinate fields $\vec{Y}(\xi^a)$ of the D-brane. The bosonic part of the low-energy effective action is the sum of a Dirac–Born–Infeld (DBI) and a Chern–Simons (CS) like term:

$$I_p = -T_p \int d^{p+1} \xi e^{-\Phi} \sqrt{-\det(\hat{G}_{ab} + \mathcal{F}_{ab})} - \rho_p \int \sum_n \hat{C}_n \wedge e^{\mathcal{F}} \quad [8]$$

where $\mathcal{F}_{ab} = \hat{B}_{ab} + 2\pi\alpha' F_{ab}$, hats denote pullbacks on the brane of bulk tensor fields (e.g., $\hat{G}_{ab} = G_{\mu\nu} \partial_a Y^\mu \partial_b Y^\nu$), F_{ab} is the field strength of the world-volume gauge field, and in the CS term one is instructed to keep the $(p + 1)$ -form of the expression under the integration sign. The constants T_p and ρ_p are the tension and charge density of the D-brane. As was the case for the effective supergravities, the above action receives curvature corrections that are higher order in the α' expansion. Note however that a class of higher-order terms have been already resummed in expression [8]. These involve arbitrary powers of F_{ab} , and are closely related more precisely T -dual, see later) to relativistic effects which can be important even in the weak-acceleration limit. When refereing to the D9 branes of the type I superstring, the action [8] includes the GS terms required to cancel the gauge anomaly.

The tension and charge density of a D p brane can be extracted from its coupling to the (closed-string) graviton and R–R $(p + 1)$ -form, with the result:

$$T_p^2 = \rho_p^2 = \frac{\pi}{\kappa^2} (4\pi^2 \alpha')^{3-p} \quad [9]$$

The equality of tension and charge follows from unbroken supersymmetry, and is also known as a Bogomol'nyi–Prasad–Sommerfeld (BPS) condition.

It implies that two or more identical D-branes exert no net static force on each other, because their R–R repulsion cancels exactly their gravitational attraction. A nontrivial check of the result [9] comes from the Dirac quantization condition (generalized to extended objects by Nepomechie and Teitelboim). Indeed, a D p brane and a D $(6 - p)$ -brane are dual excitations, like electric and magnetic charges in four dimensions, so their couplings must obey

$$2\kappa^2 \rho_p \rho_{6-p} = 2\pi k \quad \text{where } k \in \mathbb{Z} \quad [10]$$

This ensures that the Dirac singularity of the long-range R–R fields of the branes does not lead to an observable Bohm–Aharonov phase. The couplings [9] obey this condition with $k = 1$, so that D-branes carry the smallest allowed R–R charges in the theory.

A simple but important observation is that open strings living on a collection of n identical D-branes have matrix-valued wave functions ψ_{ij} , where $i, j = 1, \dots, n$ label the possible endpoints of the string. The low-energy dynamics of the branes is thus described by a nonabelian gauge theory, with group $U(n)$ if the open strings are oriented, and $SO(n)$ or $Sp(n)$ if they are not. We have already encountered such Chan–Paton factors in our discussion of the type I superstring. More generally, this simple property of D-branes has led to many insights on the geometric interpretation and engineering of gauge theories, which are reviewed in the articles Brane Construction of Gauge Theories and Gauge Theories from Strings. It has also placed on a firmer footing the idea of a brane world, according to which the fields and interactions of the standard model would be confined to a set of D-branes, while gravitons are free to propagate in the bulk (for reviews, see Brane Worlds and reference Lust (2004)). It has, finally, inspired the gauge/string theory or AdS/CFT correspondence (see AdS/CFT Correspondence and Aharony *et al.* (2000)) on which we will comment later.

Dualities and M Theory

One other key role of D-branes has been to provide evidence for the various nonperturbative duality conjectures. Dual descriptions of the same physics arise also in conventional field theory. A prime example is the Montonen–Olive duality of four-dimensional, $N = 4$ supersymmetric Yang–Mills, which is the low-energy theory describing the dynamics of a collection of D3 branes. The action

for the gauge field and six associated scalars Φ^I (all in the adjoint representations of the gauge group G) is

$$\begin{aligned} S_{N=4} = & -\frac{1}{4g^2} \int d^4x \operatorname{tr} \\ & \times \left(F_{\mu\nu} F^{\mu\nu} + 2 \sum_I D_\mu \Phi^I D^\mu \Phi^I + \sum_{I < J} 2 [\Phi^I, \Phi^J]^2 \right) \\ & - \frac{\theta}{32\pi^2} \int d^4x \operatorname{tr} (F_{\mu\nu} {}^* F^{\mu\nu}) \\ & + \text{fermionic terms} \end{aligned} \quad [11]$$

Consider for simplicity the case $G = \text{SU}(2)$. The scalar potential has flat directions along which the six Φ^I commute. By an $\text{SO}(6)$ R-symmetry rotation, we can set all but one of them to zero, and let $\langle \operatorname{tr}(\Phi^1 \Phi^1) \rangle = v^2$ in the vacuum. In this “Coulomb phase” of the theory, a $\text{U}(1)$ gauge multiplet stays massless, while the charged states become massive by the Higgs effect. The theory admits furthermore smooth magnetic-monopole and dyon solutions, and there is an elegant formula for their mass:

$$M = v |n_{\text{el}} + \tau n_{\text{mg}}|, \quad \text{where } \tau = \frac{\theta}{2\pi} + \frac{4\pi i}{g^2} \quad [12]$$

and $n_{\text{el}}(n_{\text{mg}})$ denotes the quantized electric (magnetic) charge. This is a BPS formula that receives no quantum corrections. It exhibits the $\text{SL}(2, \mathbb{Z})$ covariance of the theory,

$$\tau \rightarrow \frac{a\tau + b}{c\tau + d} \quad [13]$$

and

$$(n_{\text{el}}, n_{\text{mg}}) \rightarrow (n_{\text{el}}, n_{\text{mg}}) \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1}$$

Here a, b, c, d are integers subject to the condition $ad - bc = 1$. Of special importance is the transformation $\tau \rightarrow -1/\tau$, which exchanges electric and magnetic charges and (at least for $\theta = 0$) the strong- with the weak-coupling regimes. For more details see the review by Harvey (1996).

The extension of these ideas to string theory can be illustrated with the strong/weak-coupling duality between the type I theory, and the $\text{Spin}(32)/\mathbb{Z}_2$ heterotic string. Both have the same massless spectrum and low-energy action, whose form is dictated entirely by supersymmetry. The only difference lies in the relations between the string and supergravity parameters. Eliminating the latter, one finds

$$\lambda_{\text{het}} = \frac{1}{2\lambda_I} \quad \text{and} \quad \alpha'_{\text{het}} = \sqrt{2} \lambda_I \alpha'_I \quad [14]$$

It is thus tempting to conjecture that the strongly coupled type I theory has a dual description as a

weakly coupled heterotic string. These are, indeed, the only known ultraviolet completions of the theory [7]. Furthermore, for $\lambda_I \gg 1$, the D1 brane of the type I theory becomes light, and could be plausibly identified with the heterotic string. This conjecture has been tested successfully by comparing various supersymmetry-protected quantities (such as the tensions of BPS excitations and special higher-derivative terms in the effective action), which can be calculated exactly either semiclassically, or at a given order in the perturbative expansion. Testing the duality for nonprotected quantities is a hard and important problem, which looks currently out of reach.

The other three string theories have also well-motivated dual descriptions at strong coupling λ . The type IIB theory is believed to have an $\text{SL}(2, \mathbb{Z})$ symmetry, similar to that of the $N = 4$ super Yang-Mills. (Note that λ is a dynamical parameter, that changes with the vacuum expectation value of the dilaton $\langle \phi \rangle$. Thus, dualities are discrete gauge symmetries of string theory.) The type IIA theory has a more surprising strong-coupling limit: it grows one extra dimension (of radius $R_{11} = 1/\lambda \sqrt{\alpha'}$), and can be approximated at low energy by the maximal 11-dimensional supergravity of Cremmer, Julia, and Scherk. The latter is a very economical theory – its massless bosonic fields are only the graviton and a 3-form potential A_3 . The bosonic part of the action reads

$$\begin{aligned} S_{11D} = & \frac{1}{2\kappa_{11}^2} \int d^{11}x \sqrt{-G} (R - \tfrac{1}{2} |F_4|^2) \\ & - \frac{1}{12\kappa_{11}^2} \int A_3 \wedge F_4 \wedge F_4 \end{aligned} \quad [15]$$

The electric and magnetic charges of the 3-form are a (fundamental?) membrane and a solitonic 5-brane. Standard Kaluza-Klein reduction on a circle maps S_{11D} to the IIA supergravity action [6], where $G_{\mu\nu}$, ϕ , and C_1 descend from the 11-dimensional graviton, and B_2 and C_3 from the 3-form A_3 . Furthermore, all BPS excitations of the type IIA string theory have a counterpart in 11 dimensions, as summarized in Table 1. Finally, if one compactifies the eleventh dimension on an interval (rather than a circle), one finds the conjectured strong-coupling limit of the $E_8 \times E_8$ heterotic string.

The web of duality relations can be extended by compactifying further to $D \leq 9$ dimensions. Readers interested in more details should consult Polchinski (1998) or one of the many existing reviews of the subject (Townsend (1996), see also “Further Reading” section). In nine dimensions, in particular, the two type II theories, as well as the two heterotic superstrings, are pairwise T -dual. T -duality is a perturbative symmetry (thus firmly established, not

Table 1 BPS excitations of type IIA string theory, and their counterparts in \mathcal{M} theory compactified on a circle of radius R_{11}

Tension	Type IIA	\mathcal{M} on S^1	Tension
$(\sqrt{\pi}/\kappa_{10})(2\pi\sqrt{\alpha'})^3$	D0 brane	K–K excitation	$1/R_{11}$
$T_F = (2\pi\alpha')^{-1}$	String	Wrapped membrane	$2\pi R_{11}(2\pi^2/\kappa_{11}^2)^{1/3}$
$(\sqrt{\pi}/\kappa_{10})(2\pi\sqrt{\alpha'})$	D2 brane	Membrane	$T_2^M = (2\pi^2/\kappa_{11}^2)^{1/3}$
$(\sqrt{\pi}/\kappa_{10})(2\pi\sqrt{\alpha'})^{-1}$	D4 brane	Wrapped 5-brane	$R_{11}(2\pi^2/\kappa_{11}^2)^{2/3}$
$(\pi/\kappa_{10}^2)(2\pi\alpha')$	NS-5-brane	5-brane	$(1/2\pi)(2\pi^2/\kappa_{11}^2)^{2/3}$
$(\sqrt{\pi}/\kappa_{10})(2\pi\sqrt{\alpha'})^{-3}$	D6 brane	K–K monopole	$2\pi^2 R_{11}^2/\kappa_{11}^2$

From Bachas CP (1997) Lectures on D-branes. In: Olive DI and West PC (eds.) *Duality and Supersymmetric Theories*, Proceedings, Easter School, Newton Institute, Euroconference, Cambridge, UK, April 7–18. With permission of Cambridge University Press.

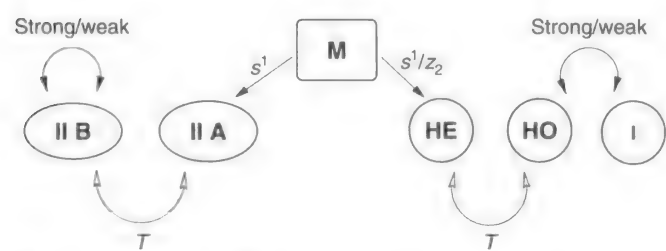


Figure 4 Web of dualities in nine dimensions. From Bachas CP (1997) Lectures on D-branes. In: Olive DI and West PC (eds.) *Duality and Supersymmetric Theories*, Proceedings, Easter School, Newton Institute, Euroconference, Cambridge, UK, April 7–18. With permission of Cambridge University Press.

only conjectured) which exchanges momentum and winding modes. Putting together all the links one arrives at the fully connected web of Figure 4. This makes the point that all five consistent superstrings, and also 11-dimensional supergravity, are limits of a unique underlying structure called M theory. (For lack of a better definition, “M” is sometimes also used to denote the $D=11$ supergravity plus supermembranes, as in Figure 4.) A background-independent definition of M theory has remained elusive. Attempts to define it as a matrix model of D0 branes, or by quantizing a fundamental membrane, proved interesting but incomplete. A difficulty stems from the fact that in a generic background, or in $D=11$ Minkowski spacetime, there is only a dimensionful parameter fixing the scale at which the theory becomes strongly coupled.

Other Developments and Outlook

We have not discussed in this brief review some important developments covered in other contributions to the encyclopedia. For the reader’s convenience, and for completeness, we enumerate (some of) them giving the appropriate cross-references:

Compactification. To make contact with the standard model of particle physics, one has to

compactify string theory on a six-dimensional manifold. There is an embarrassment of riches, but no completely realistic vacuum and, more significantly, no guiding dynamical principle to help us decide (see Compactification of Superstring Theory). The controlled (and phenomenologically required) breaking of spacetime supersymmetry is also a problem.

Conformal field theory and quantum geometry. The algebraic tools of 2D conformal field theory, both bulk and boundary (see Two-Dimensional Conformal Field Theory and Vertex Operator Algebras), play an important role in string theory. They allow, in certain cases, a resummation of α' effects, thereby probing the regime where classical geometric notions do not apply.

Microscopic models of black holes. Charged extremal black holes can be modeled in string theory by BPS configurations of D-branes. This has led to the first microscopic derivation of the Bekenstein–Hawking entropy formula, a result expected from any consistent theory of quantum gravity. As with the tests of duality, the extension of these results to neutral black holes is a difficult open problem – see Branes and Black Hole Statistical Mechanics.

AdS/CFT and holography. A new type of (holographic) duality is the one that relates supersymmetric gauge theories in four dimensions to string theory in asymptotically anti-de Sitter spacetimes. The sharpest and best-tested version of this duality relates $N=4$ super Yang–Mills to string theory in $AdS_5 \times S_5$. Solving the σ -model in this latter background is one of the keys to further progress in the subject (see AdS/CFT Correspondence).

String phenomenology. Finding an experimental confirmation of string theory is clearly one of the most pressing outstanding questions. There exist several interesting possibilities for this – cosmic strings, large extra dimensions, modifications of gravity, primordial cosmology (see String Theory: Phenomenology for a

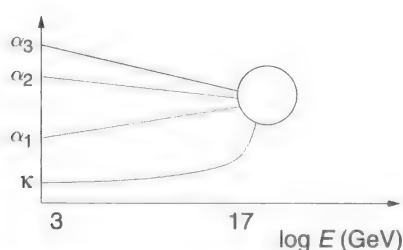


Figure 5 The unification of couplings.

review). Here we point out the one supporting piece of experimental evidence: the unification of the gauge couplings of the (supersymmetric, minimal) standard model at a scale close to, but below the Planck scale, as illustrated in **Figure 5**. This is a generic “prediction” of string theory, especially in its heterotic version.

See also: AdS/CFT Correspondence; Boundary Conformal Field Theory; Brane Construction of Gauge Theories; Brane Worlds; Branes and Black Hole Statistical Mechanics; Compactification of Superstring Theory; Derived Categories; Electroweak Theory; Gauge Theories from Strings; Noncommutative Geometry from Strings; Supermanifolds; String Field Theory; String Theory: Phenomenology; Supergravity; Two-Dimensional Conformal Field Theory and Vertex Operator Algebras; Wheeler–DeWitt Theory.

Further Reading

Aharony O, Gubser SS, Maldacena JM, Ooguri H, and Oz Y (2000) Large N field theories, string theory and gravity. *Physics Report* 323: 183 (arXiv:hep-th/9905111).

Angelantonj C and Sagnotti A (2002) Open strings. *Physics Report* 371: 1.

Angelantonj C and Sagnotti A (2003) Open strings – erratum. *Physics Report* 376: 339 (arXiv:hep-th/0204089).

Bachas CP (1997) Lectures on D-branes. In: Olive DI and West PC (eds.) *Duality and Supersymmetric Theories*, Proceedings, Easter School, Newton Institute, Euroconference, Cambridge, UK, April 7–18, arXiv:hep-th/9806199.

Berkovits N (2002) ICTP lectures on covariant quantization of the superstring, arXiv:hep-th/0209059.

D'Hoker E and Phong DH (2002) Lectures on two-loop superstrings, arXiv:hep-th/0211111.

Green MB, Schwarz JH, and Witten E (1987) *Superstring Theory*. Vol. 1: *Introduction*; Vol. 2: *Loop Amplitudes, Anomalies and Phenomenology*. Cambridge: Cambridge University Press.

Harvey JA (1996) Magnetic monopoles, duality, and supersymmetry. arXiv:hep-th/9603086.

Kiritis E (2004) D-branes in standard model building, gravity and cosmology. *Fortschritte der Physik* 52: 200 (arXiv:hep-th/0310001).

Lust D (2004) Intersecting brane worlds: A path to the standard model? *Classical and Quantum Gravity* 21: S1399 (arXiv:hep-th/0401156).

Obers NA and Pioline B (1999) U-duality and M-theory. *Physics Report* 318: 113 (arXiv:hep-th/9809039).

Polchinski J (1998) *String Theory*. Vol. 1: *An Introduction to the Bosonic String*; Vol. 2: *Superstring Theory and Beyond*. Cambridge: Cambridge University Press.

Salam A and Sezgin E (1989) *Supergravities in Diverse Dimensions*. Amsterdam: North-Holland/World Scientific.

Sen A (1997) An introduction to non-perturbative string theory, arXiv:hep-th/9802051.

The following URL of the HEP-SPIRES database: <http://www.slac-stanford.edu> collects many popular string theory reviews.

Townsend PK (1996) Four lectures on M-theory, arXiv:hep-th/9612121.

Zwiebach B (2004) *A First Course in String Theory*. Cambridge: Cambridge University Press.

Supersymmetric Particle Models

S Pokorski, Warsaw University, Warsaw, Poland

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Supersymmetric quantum field theories (see Supergravity) are characterized by the existence of one ($N=1$ supersymmetry) or several ($N>1$ extended supersymmetry) conserved Noether-like charges Q_A , $A=1, \dots, N$, which establish symmetry links between particle states of different spin. Supersymmetry ensures equal numbers of bosonic and fermionic particle states. If it is exact, bosons and fermions related by supersymmetry transformations have equal masses. Moreover, supersymmetry

imposes stringent relations between interactions which involve particles of different spin. This gives rise to a special ultraviolet behavior of supersymmetric theories. Their ultraviolet divergences are much softer than in nonsupersymmetric theories. In particular, $N=4$ supersymmetric quantum field theories are finite and for any N they are free from quadratic divergences plaguing ordinary theories with elementary scalars. $N>4$ supersymmetric theories necessarily involve particles of spin higher than 1 and are not renormalizable. Supersymmetry promoted to a local symmetry includes gravity.

Only $N=1$ supersymmetric theories allow for chiral fermions which are the fundamental objects in elementary particle interactions (see Standard Model of Particle Physics). This is because parity and

charge conjugation symmetries are violated in weak interactions. Therefore, $N > 1$ theories may not be of immediate phenomenological relevance. However, they may be useful for constructing supersymmetric theories in more than four dimensions (more than three spatial dimensions). Chiral (effective) theory in four dimensions can be then obtained after compactification of extra dimensions. For instance, $N=2$ theory in five dimensions (x_μ, y) compactified on a circle with reflection symmetry $y \rightarrow -y$ (orbifold compactification) gives chiral $N=1$ theory in four dimensions.

Absence of quadratic divergences in supersymmetric theories is the main argument supporting the belief that fundamental interactions of elementary particles at energies not higher than $\mathcal{O}(1 \text{ TeV})$ should be described by an (approximately) $N=1$ supersymmetric extension of the standard model (SM). Indeed, supersymmetric models elegantly solve the so-called hierarchy problem of the SM. At present, supersymmetry remains a theoretical hypothesis. No experimental evidence for it has been found yet (for experimental lower bounds on the masses of supersymmetric particles see Eidelman *et al.* (2004)). Supersymmetric models will be tested experimentally at the Large Linear Collider at CERN (Geneva), after the completion of its construction in 2007. Supergravity theories may be physically relevant as an intermediate step in constructing phenomenologically viable models from superstring theories.

The essence of the hierarchy problem of the standard model (SM) – the successful $SU(3)_c \times SU(2)_L \times U(1)_Y$ gauge theory of interactions of quarks and leptons at energies up to about 100 GeV – is the following. By itself, the SM does not explain the value of the Fermi scale ν of the electroweak $SU(2)_L \times U(1)_Y$ symmetry breaking ($\nu \sim G_F^{-1/2}$ where G_F is the Fermi constant determined by the life time of the muon). Indeed, in the SM, the electroweak symmetry breaking is realized by an elementary Higgs field H (an $SU(2)$ doublet) with a potential

$$V = m^2 H^\dagger H + \frac{\lambda}{2} (H^\dagger H)^2 \quad [1]$$

where m and λ are free parameters of the SM. When $m^2 < 0$ is chosen, the minimum of the potential occurs when

$$\langle H^\dagger H \rangle = -\frac{m^2}{\lambda} \equiv \frac{\nu^2}{2} \quad [2]$$

that is, the Higgs doublet acquires $SU(2) \times U(1)_Y$ breaking vacuum expectation value ν which is just the Fermi scale. The masses of the intermediate vector bosons W^\pm and Z^0 are proportional to ν and depend also on the gauge couplings. Within the SM

understood as a theory with the momentum cut-off Λ_{SM} , quantum corrections to the mass parameter m^2 in eqn [1] are quadratically divergent:

$$\delta m^2 = \frac{3}{64\pi^2} (3g_2^2 + g_1^2 + \lambda - 8y_t^2) \Lambda_{SM}^2 + \dots \quad [3]$$

Here, g_1, g_2 , and y_t are the gauge couplings of the groups $U(1)_Y, SU(2)_L$, and the top-quark Yukawa coupling, respectively. This means that if, above the energy scale Λ_{SM} , the SM is replaced by some more fundamental theory, in which there are particles of masses $M \gtrsim \Lambda_{SM}$, the quantum corrections to m^2 are quadratically dependent on the new mass scale M . For $M \gg \nu$, this is very unnatural even if the original parameter m^2 remains a free parameter of this underlying theory and particularly difficult to accept if in the underlying theory m^2 is fixed by some more fundamental considerations. If the SM was the correct theory up to, for example, the mass scale suggested by the see-saw mechanism for the neutrino masses, $\Lambda_{SM} \sim 10^{15} \text{ GeV}$

$$|\delta m^2| \sim 10^{28} \text{ GeV}^2 \sim 10^{24} \nu^2!$$

Clearly, this excludes the possibility of understanding the magnitude of the Fermi scale ν in any sensible way. Thus, for naturalness of the Higgs mechanism in the SM there should exist a new mass scale $M \gtrsim \nu$, say only one order of magnitude higher than ν and the theory describing the physics above that scale should be free of quadratic divergences. (Approximate) supersymmetry is at present the most elegant and theoretically most complete solution to the hierarchy problem of the SM.

Supersymmetric Extensions of the SM

In supersymmetry, the gauge fields A_μ^a are promoted to vector superfields $\hat{V}^a = (A_\mu^a, \lambda^a, D^a)$, one for each gauge symmetry group generator, where λ^a 's are Weyl fermions (called gauginos) and D^a 's are nondynamical auxiliary fields. A renormalizable supersymmetric gauge theory is completely defined (see, e.g., Sohnius (1985) and Wess and Bagger (1992)) by specifying the gauge group, the set of chiral supermultiplets $\hat{\Phi}_i = (\phi_i, \psi_i, F_i)$ representing matter fields, and the superpotential – a holomorphic polynomial function of at most third order in the chiral superfields which determines Yukawa couplings of the fermions ψ_i and scalars ϕ_i . Auxiliary fields D^a and F_i can be eliminated via their (algebraic) equations of motion.

The so-called minimal supersymmetric SM (MSSM) encodes the main features of any supersymmetric extension of the SM. Its gauge group is

$SU(3) \times SU(2) \times U(1)$ – the same as in the SM – and the chiral superfields are associated to each of the SM quark and lepton fields. Thus, quarks and leptons get scalar spin zero superpartners, the squarks and sleptons, carrying the same quantum numbers as their corresponding fermions and the vector superfields provide spin 1/2 superpartners for the gauge fields – the gluinos, the winos, and the bino. The SM Higgs doublet with weak hypercharge $Y=1/2$ becomes a scalar component of a chiral superfield \hat{H}_u which contains in addition one doublet of Weyl fermions – the Higgsinos. The chiral anomaly cancelation condition requires that there be also a second Higgs chiral superfield \hat{H}_d with $Y=-1/2$. Such a superfield is also required for giving masses to all flavors of quarks; because of the holomorphicity of the superpotential the same Higgs doublet cannot couple simultaneously to all quarks.

With the MSSM superfield content, the most general renormalizable superpotential consistent with the gauge symmetry has the form

$$W = Y_u \hat{U}^c \hat{Q} \hat{H}_u + Y_d \hat{D}^c \hat{Q} \hat{H}_d + Y_l \hat{E}^c \hat{L} \hat{H}_d + \mu \hat{H}_d \hat{H}_u \\ + \lambda_1 \hat{D}^c \hat{Q} \hat{L} + \lambda_2 \hat{E}^c \hat{L} \hat{L} + \lambda_3 \hat{U}^c \hat{D}^c \hat{D}^c + \lambda_4 \hat{L} \hat{H}_u \quad [4]$$

(flavor indices are suppressed) where the superfield \hat{Q} contains the $SU(2)$ quark doublet Q and its scalar superpartner \tilde{Q} and similarly for the lepton doublet \hat{L} , quark singlets \hat{U}, \hat{D} , and lepton singlet \hat{E} superfields. The three first terms in [4] give the SM-like Yukawa couplings of quarks and leptons to the Higgs fields together with Yukawa couplings of the corresponding superpartners. The fourth term has no SM analogy; it gives supersymmetric masses to the Higgs scalar and Higgsinos. The interactions in the second line do not conserve baryon and lepton numbers, respectively B and L , and should be forbidden (or strongly suppressed) by some additional symmetry of the theory as they would lead to rapid proton decay. A discrete symmetry, called R -parity $R=(-1)^{2S+3(B-L)}$, where S is the spin of the field, is an interesting possibility. R -parity acts differently on the different components of the superfields: it is even for all SM particles and odd for their superpartners. Its conservation implies that superpartners must appear in pairs in any interaction vertex. Thus, with R -parity imposed, the lightest supersymmetric particle is stable and it is an excellent candidate for the dark matter in the universe.

Supersymmetry cannot be an exact symmetry of nature because there do not exist elementary fermions and bosons degenerate in mass. The superpotential [4] does not break supersymmetry spontaneously but even if it did the elementary fermions and bosons would on average have equal masses (they would satisfy some mass sum rule) which is also

contradicted by the experimental data. Therefore, in the MSSM, supersymmetry has to be broken explicitly but in such a way that the soft ultraviolet behavior remains intact. Remarkably, the supersymmetry breaking terms which can be added to the MSSM Lagrangian without reintroducing quadratic divergences make heavy just those fields which are opposite statistics superpartners of the SM gauge bosons and fermions. These so-called soft terms are:

$$\mathcal{L}_{\text{soft}} = -\frac{1}{2}\tilde{G}\tilde{G}^a\tilde{G}^a - \frac{1}{2}\tilde{W}\tilde{W}^a\tilde{W}^a - \frac{1}{2}\tilde{B}\tilde{B}\tilde{B} \\ - m_Q^2|\tilde{Q}|^2 - m_U^2|\tilde{U}^c|^2 - m_D^2|\tilde{D}^c|^2 \\ - m_L^2|\tilde{L}|^2 - m_E^2|\tilde{E}^c|^2 - m_{H_u}^2|H^u|^2 \\ - m_{H_d}^2|H^d|^2 - m_3^2(H^u H^d + \text{c.c.}) \\ + A_U \tilde{U}^c \tilde{Q} H_u + A_D \tilde{D}^c \tilde{Q} H_d + A_E \tilde{E}^c \tilde{L} H_d \quad [5]$$

and yield gaugino (gluino \tilde{G} , wino \tilde{W} , and bino \tilde{B}) and scalar mass terms as well as explicit trilinear couplings between scalars (scalar mass terms and A -terms are 3×3 matrices in the flavor space). As a result, supersymmetry is broken in the mass spectra but not in the dimensionless couplings.

The origin of the soft supersymmetry breaking remains an open issue. Terms [5] are most probably remnants of the spontaneous supersymmetry breaking in the so-called “hidden” sector – a hypothetical set of fields that do not interact directly with the MSSM fields. For example, in the popular scenario, they interact with the MSSM fields only gravitationally and spontaneous supersymmetry breaking in the hidden sector is communicated to the MSSM sector by gravitational interactions giving rise to terms [5]. Several other mechanisms of supersymmetry breaking transmission have also been proposed (gauge mediation, anomaly mediation, etc.).

The mass parameters and A -terms in [5] are free parameters of the low-energy supersymmetric theory and, combined with the interactions like $Q\tilde{Q}\tilde{G}$ originating from supersymmetric kinetic terms, may be a new, troublesome, source of flavor changing neutral currents and of CP violation.

Higgs Sector of the MSSM

The MSSM Higgs potential reads

$$V = m_1^2|H_d|^2 + m_2^2|H_u|^2 + m_3^2(H_u H_d + \text{c.c.}) \\ + \frac{g_1^2 + g_2^2}{8} (|H_d|^2 - |H_u|^2)^2 \quad [6]$$

Its quartic part is uniquely determined by the structure of the supersymmetric gauge theory. The parameters m_1^2 , m_2^2 , and m_3^2 are determined by

the soft supersymmetry breaking Higgs boson masses [5] and the μ parameter in [4]. The potential [6] is bounded from below for $m_1^2 + m_2^2 > 2m_3^2$, and for $m_1^2 m_2^2 - m_3^2 < 0$ it has the electroweak symmetry breaking minimum at $v_u = \langle H_u^0 \rangle \neq 0, v_d = \langle H_d^0 \rangle \neq 0$. The ratio $v_u/v_d \equiv \tan \beta$ is then phenomenologically a very important parameter.

Quantum corrections to the mass parameters in [6] are controlled by the mass scale M_{soft} of the supersymmetry breaking terms [5]; at the one-loop level instead of [3], one finds

$$\delta m_{1,2}^2 \sim \frac{1}{16\pi^2} (3g_2^2 + g_1^2 - 12y_{b,t}^2) M_{\text{soft}}^2 \ln \frac{\Lambda_{\text{NEW}}^2}{M_{\text{soft}}^2} \quad [7]$$

where y_b and y_t are the bottom- and top-quark Yukawa couplings, respectively and Λ_{NEW} is the scale at which the soft supersymmetry breaking terms are generated by the putative supersymmetry breaking transmission mechanism. In gravity mediation scenarios, $\Lambda_{\text{NEW}} \sim M_{\text{Pl}}$. In gauge mediation scenarios, Λ_{NEW} is low but it is a new scale, introduced by hand.

In the softly broken supersymmetric models, the hierarchy problem is solved for $M_{\text{soft}} \lesssim \mathcal{O}(10)\nu$. Moreover, eqn [7] shows that via quantum corrections the large top-quark Yukawa coupling y_t drives the mass parameter m_2^2 to a negative value, inducing the electroweak symmetry breaking. This means that in supersymmetric models the electroweak scale is calculable in terms of the known coupling constants and the (unknown) scales M_{soft} and cutoff scale Λ_{NEW} to the MSSM. If $M_{\text{soft}} \lesssim \mathcal{O}(10)\nu$, the correct electroweak scale is obtained for $\Lambda_{\text{NEW}} \sim M_{\text{GUT}}$. This nicely fits with unification of the gauge couplings.

In supersymmetric models, the quartic couplings in the Higgs potential are restricted. This typically leads to a strong upper bound on the mass of the lightest Higgs particle. In the minimal model with the potential [6], at the tree level

$$M_{\text{Higgs}} < M_Z \approx 91 \text{ GeV} \quad [8]$$

This bound is substantially modified by quantum corrections. They depend quadratically on the top-quark mass and logarithmically on the stop mass scale $M_{\tilde{t}} \sim M_{\text{soft}}$:

$$M_{\text{Higgs}}^2 < \lambda v^2 \quad [9]$$

where λ is given by

$$\lambda = \frac{1}{8}(g_2^2 + g_1^2) \cos^2 2\beta + \Delta\lambda$$

$$\text{with } \Delta\lambda = \frac{3g_2^2}{8\pi^2} \frac{m_t^4}{v^2 M_W^2} \ln \frac{M_{\tilde{t}}^2}{m_t^2} \quad [10]$$

For $M_{\tilde{t}} < 1 \text{ TeV}$, $M_{\text{Higgs}} < 130 \text{ GeV}$.

The minimal-model bound on the Higgs mass can be relaxed in models with extended Higgs sector. For instance, if an additional gauge group singlet chiral superfield couples to the Higgs doublets, the Higgs self-coupling λ in [9] receives additional contributions. Explicit calculations show that in such and other models, with $M_{\text{soft}} \lesssim 1 \text{ TeV}$, the bound on the Higgs mass cannot be raised above $\sim 150 \text{ GeV}$ if one wants to preserve perturbative gauge coupling unification.

Supersymmetric Grand Unified Theories

There are two striking aspects of the matter spectrum in the SM. One is the chiral anomalies cancelation (Weinberg 1996–2000, Pokorski 2000), which is necessary for a unitary (and renormalizable) theory, and occurs thanks to certain conspiracy between quarks and leptons suggesting a deeper link between them. The second one is that the spectrum fits into simple representations of the SU(5) and SO(10) groups (Ross 1985). Indeed, each generation of the SM matter fills $5^* + 10 + 1$ (if the right-handed neutrino is included into the spectrum) representations of SU(5) and for SO(10), $16 = 5^* + 10 + 1$. The assignment of fermions to the SU(5) or SO(10) representations fixes the normalization of the U(1)_Y generator. Both facts suggest unification of strong and electroweak elementary forces in a grand unified theory with some bigger gauge symmetry group. Such unification implies that all the SM gauge forces become of equal strength at some unification scale. Their strength is measured by the running gauge couplings $\alpha_i = g_i^2/4\pi$, $i = 1, 2, 3$, of the three group factors SU(3)_c \times SU(2)_L \times U(1)_Y. The energy scale dependence of α_i is governed by the renormalization group equations. In the first nontrivial approximation, they read:

$$\frac{1}{\alpha_i(Q)} = \frac{1}{\alpha_i(M_Z)} - \frac{b_0^{(i)}}{2\pi} \ln \left(\frac{Q}{M_Z} \right) \quad [11]$$

Here, $1/\alpha_i(M_Z) = (58.98 \pm 0.04, 29.57 \pm 0.03, 8.40 \pm 0.14)$ are the experimental values of the gauge couplings at the Fermi scale and $b_0^{(i)}$ are the coefficients which depend on the matter content of the theory. They are

$$b_0 = \left(\frac{1}{10} + \frac{4}{3}N_g, -\frac{43}{6} + \frac{4}{3}N_g, -11 + \frac{4}{3}N_g \right)$$

in the SM and

$$b_0 = \left(\frac{3}{5} + 2N_g, -5 + 2N_g, -9 + 2N_g \right)$$

in the MSSM, where N_g is the number of fermion generations. In the SM, the running gauge couplings

approach each other at high scale of order 10^{13} GeV but never unify.

In the MSSM, with sparticle spectrum characterized by $M_{\text{soft}} \approx 1$ TeV and for the initial Fermi scale values given above, the three gauge couplings unify with high precision at the scale $M_{\text{GUT}} \sim 10^{16}$ GeV. Therefore, the MSSM can be embedded into supersymmetric grand unified theories with no hierarchy problem for the Fermi scale (it is stable with respect to radiative corrections generated by particles with masses $\sim M_{\text{GUT}}$) and no conflict with the measured values of the gauge couplings.

In the SM, the baryon number is (perturbatively) conserved since there are no renormalizable couplings violating this symmetry. Experimental search for proton decay, for example, $p \rightarrow e^+ \pi^0$, $p \rightarrow K^+ \nu$, is one of the most fundamental tests for particle physics. The present limit on the proton life time is $\tau_p > 10^{33}$ yr. In grand unified theories, baryon number conservation is violated by interactions mediated by the heavy gauge bosons corresponding to the enlarged gauge symmetry (e.g., SU(5)), spontaneously broken at M_{GUT} to the SM gauge symmetry. Such interactions manifest themselves at low energy as additional, nonrenormalizable interactions added to the SM Lagrangian. Proton decay is then induced by the set of dimension-6 operators of the form

$$O_i^{(6)} = \frac{c_i^{(6)}}{M_{(6)}^2} qqql \quad [12]$$

where q, l denote quarks and leptons, respectively. For $c_i^{(6)} \sim \alpha_{\text{GUT}} \approx 1/25$, the experimental limit on τ_p requires $M_{(6)} \gtrsim 10^{15}$ GeV, consistently with $M_{\text{GUT}} = 10^{16}$ GeV in supersymmetric GUTs. However, in supersymmetric GUTs, there is still another, genuinely supersymmetric, source of contributions to the proton decay amplitudes. These are the dimension-5 operators

$$O_i^{(5)} = \frac{c_i^{(5)}}{M_{(5)}^2} qq\tilde{q}\tilde{l} \quad [13]$$

where \tilde{q}, \tilde{l} denote squarks and sleptons, respectively. Such operators originate from the exchange of the color triplet scalars present in the Higgs boson GUT multiplets, with $M_{(5)} \sim M_{\text{GUT}} \sim 10^{16}$ GeV, and $c^{(5)} \gtrsim 10^{-7}$ is given by the Yukawa couplings. Inserted into diagrams with gaugino exchanges they give rise to dimension-6 operators of the form [12]. One then gets $c^{(6)} = \alpha_{\text{GUT}} c^{(5)}$, $M_{(6)}^2 = M_{(5)} M_{\text{SUSY}}$. Given various uncertainties, for example, in the unknown squark, gaugino, and heavy Higgs boson mass spectrum, such contributions in supersymmetric GUT models predict the proton life time to

be consistent with but close to the present experimental limits.

Summary

Supersymmetry is distinct in several very important points from all other proposed solutions to the hierarchy problem. First of all, it provides a general theoretical framework which allows one to address many physical questions. Supersymmetric models, like the MSSM or its simple extensions, satisfy a very important criterion of “perturbative calculability.” In particular, they are easily consistent with the precision electroweak data. The SM is their low-energy approximation in the sense of the Appelquist–Carazzone decoupling, so most of the successful structure of the SM is built into supersymmetric models. The quadratically divergent quantum corrections to the Higgs mass parameter (the origin of the hierarchy problem in the SM) are absent in any order of perturbation theory. Therefore, the cutoff to a supersymmetric theory can be as high as the Planck scale, and “small” scale of the electroweak breaking is still natural. Supersymmetry is not only consistent with grand unification of elementary forces but, in fact, makes it very successful. And, finally, supersymmetry is needed for string theory.

However, there are also some problems to be solved: the hierarchy problem of the electroweak scale is solved but the origin of the soft supersymmetry breaking scale M_{soft} remains an open question: spontaneous supersymmetry breaking and its transmission to the visible sector is a difficult problem and a fully satisfactory mechanism which would yield M_{soft} hierarchically smaller than the Planck (string) scale has not yet been found. On the phenomenological side, there are new potential sources of flavor-changing neutral current transitions and of CP violation, and baryon and lepton numbers are not automatically conserved by the renormalizable couplings. But even those problems can at least be discussed in a concrete quantitative way.

See also: Brane Construction of Gauge Theories; Perturbation Theory and its Techniques; Seiberg–Witten Theory; Standard Model of Particle Physics; Supergravity; Supermanifolds.

Further Reading

- Eidelman S *et al.* (The Particle Data Group) (2004) Review of particle physics. *Physics Letters B* 592: 1.
- Kane GL (ed.) (1998a) *Perspectives in Supersymmetry*. Singapore: World Scientific.
- Kane GL (ed.) (1998b) *Perspectives on Higgs Physics II*. Singapore: World Scientific.
- Nilles H-P (1984) *Physics Reports C* 110: 1.

- Pokorski S (2000) *Gauge Field Theories*, Cambridge Monographs on Mathematical Physics, 2nd edn. Cambridge: Cambridge University Press.
- Ross GG (1985) *Grand Unified Theories*. Redwood City, CA: Addison-Wesley.
- Sohnius M-F (1985) *Physics Reports C* 128: 39.

- Weinberg S (1996–2000) *The Quantum Theory of Fields*, vols. I–III. Cambridge: Cambridge University Press.
- Wess J and Bagger J (1992) *Supersymmetry and Supergravity*, Princeton Series in Physics, 2nd edn. Princeton: Princeton University Press.

Supersymmetric Quantum Mechanics

J-W van Holten, NIKHEF, Amsterdam, The Netherlands

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Supersymmetric quantum mechanics is a specific extension of quantum mechanics with fermionic degrees of freedom. In quantum field theory and many-body theory, a fermionic degree of freedom is one which is subject to Pauli's principle: any nondegenerate quantum state associated with a fermionic degree of freedom can be occupied at most once at any time. Similarly, in quantum mechanics, one associates a fermionic degree of freedom with an observable, the eigenvalue spectrum of which is restricted to the discrete set $(0, 1)$.

The simplest example of a purely fermionic quantum system is the fermionic oscillator. It is represented by conjugate operators (f, f^\dagger) such that

$$f^2 = 0, \quad f^{\dagger 2} = 0, \quad ff^\dagger + f^\dagger f = 1 \quad [1]$$

with a Hamiltonian H given by the bilinear expression

$$H_f = \varepsilon_f + \hbar\omega f^\dagger f \quad [2]$$

The state space of this system is spanned by two independent state vectors $|0\rangle$ and $|1\rangle$, such that

$$\begin{aligned} f|0\rangle &= 0, & f^\dagger|0\rangle &= |1\rangle \\ f|1\rangle &= |0\rangle, & f^\dagger|1\rangle &= 0 \end{aligned} \quad [3]$$

By construction, the states $|n_f\rangle$ are eigenstates of fermion number,

$$N_f = f^\dagger f, \quad N_f^2 = N_f \quad [4]$$

with eigenvalue $n_f = (0, 1)$; this implements the Pauli principle. The states have energy eigenvalues

$$E_{n_f} = \varepsilon_f + n_f \hbar\omega, \quad n_f = (0, 1) \quad [5]$$

differing in energy by $\Delta E = \hbar\omega$. Physically, the system can be identified with a single fixed magnetic dipole in

an external magnetic field, the only polarization states of the dipole being spin up or spin down.

In the Schrödinger representation of quantum mechanics (wave mechanics), fermionic degrees of freedom are represented by anticommuting Grassmann variables. These have no immediate classical analog, but can be used to construct quasiclassical observables like spin.

A supersymmetric quantum system is a system possessing both fermionic and bosonic degrees of freedom, characterized by a degeneracy between states with even and odd fermion number. In the Schrödinger representation, this is manifest in a symmetry transforming bosonic (Grassmann-even) into fermionic (Grassmann-odd) variables. The generators of the supersymmetry transformations square to the Hamiltonian of the system.

The Supersymmetric Oscillator

An elementary example of a supersymmetric quantum system is the supersymmetric oscillator. It is a physical system combining a standard bosonic quantum oscillator with a fermionic oscillator of the same frequency. The ordinary harmonic oscillator is described by the pair of lowering and raising operators (b, b^\dagger) , with commutator

$$bb^\dagger - b^\dagger b = 1 \quad [6]$$

and the Hamiltonian

$$H_b = \varepsilon_b + \hbar\omega b^\dagger b \quad [7]$$

In this case, the eigenvalue spectrum of the occupation number

$$N_b = b^\dagger b \quad [8]$$

consists of all non-negative integers $n_b = 0, 1, 2, \dots$, with corresponding energy eigenvalues. To construct the supersymmetric oscillator, the harmonic oscillator is combined with a fermionic oscillator [2] of the same frequency:

$$H_s = \varepsilon_0 + \hbar\omega(b^\dagger b + f^\dagger f) \quad [9]$$

where $\varepsilon_0 = \varepsilon_b + \varepsilon_f$. The ground state of this system is the state annihilated by both b and f :

$$b|0,0\rangle = f|0,0\rangle = 0 \quad [10]$$

The full set of energy eigenstates of the system is constructed by taking

$$|n_b, n_f\rangle = \frac{1}{\sqrt{n_b!}} b^{\dagger n_b} f^{\dagger n_f} |0,0\rangle \quad [11]$$

$$n_b = (0, 1, 2, \dots), \quad n_f = (0, 1)$$

with the energy eigenvalue spectrum

$$E(n_b, n_f) = \varepsilon_0 + n\hbar\omega, \quad n = n_b + n_f \quad [12]$$

Clearly, there is a degeneracy in energy between the states $|n_b + 1, 0\rangle$ and $|n_b, 1\rangle$, which have the same total occupation number n , but differ in the bosonic and fermionic occupation number by one unit. This is illustrated in **Figure 1**. Such pairs of states which are degenerate in energy can be transformed into each other by the operators

$$Q = \sqrt{2\hbar\omega} b^\dagger f, \quad Q^\dagger = \sqrt{2\hbar\omega} f^\dagger b \quad [13]$$

The explicit transformations are

$$|n_b + 1, 0\rangle = \frac{1}{\sqrt{2(n_b + 1)\hbar\omega}} Q |n_b, 1\rangle \quad [14]$$

$$|n_b, 1\rangle = \frac{1}{\sqrt{2(n_b + 1)\hbar\omega}} Q^\dagger |n_b + 1, 0\rangle$$

The operations [14] are called supersymmetry transformations, and the operators Q and Q^\dagger are called supercharges.

As the zero point of energy is arbitrary in systems without gravitational interactions, it is customary to take $\varepsilon_0 = 0$, that is, $\varepsilon_f = -\varepsilon_b$; with the normalization [13], the Hamiltonian H is then the symmetrized absolute square of the supercharges:

$$QQ^\dagger + Q^\dagger Q = 2H \quad [15]$$

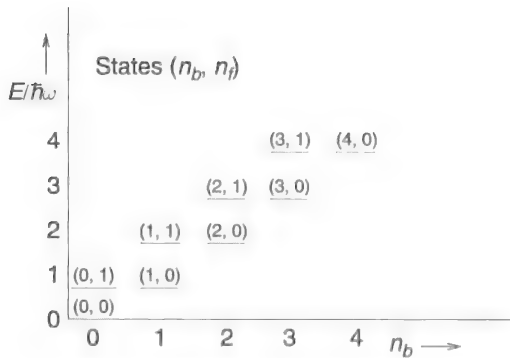


Figure 1 Spectrum of states of the supersymmetric oscillator.

whilst

$$Q^2 = Q^{\dagger 2} = 0 \quad [16]$$

The above relations suffice to guarantee that the supercharges (Q, Q^\dagger) are conserved:

$$[Q, H] = [Q^\dagger, H] = 0 \quad [17]$$

a result re-expressing the degeneracy between states with the same n but different n_b and n_f . The real form of the supercharges is

$$Q_1 = \frac{1}{2}(Q + Q^\dagger), \quad Q_2 = \frac{1}{2i}(Q - Q^\dagger) \quad [18]$$

In this representation

$$H = Q_1^2 + Q_2^2 \quad [19]$$

An important observation is that the ground state is the only state annihilated by both supersymmetry operators:

$$Q|0,0\rangle = 0, \quad Q^\dagger|0,0\rangle = 0 \quad [20]$$

Indeed, it is the only state with zero energy eigenvalue, and only such a state can be an invariant supersinglet; all other states have positive energy and they necessarily occur in supersymmetry pairs.

Anticommuting Variables

Fermionic degrees of freedom can be described in a pseudoclassical formulation by anticommuting variables ξ taking values in an infinite-dimensional Grassmann algebra:

$$\xi\xi' + \xi'\xi = 0 \quad [21]$$

With an anticommuting variable ξ , we can associate a derivative operator $\partial/\partial\xi$, which is an element of another Grassmann algebra such that

$$\left\{ \frac{\partial}{\partial\xi}, \xi \right\} = \frac{\partial}{\partial\xi} \xi + \xi \frac{\partial}{\partial\xi} = 1, \quad \frac{\partial^2}{\partial\xi^2} = 0 \quad [22]$$

This extends the original Grassmann algebra to a Clifford algebra. Integration with respect to an anticommuting variable is defined in the same way:

$$\int d\xi \cdot \xi = 1, \quad \int d\xi \cdot 1 = 0 \quad [23]$$

that is, integration is the same as differentiation for anticommuting variables. With these definitions, we can represent the fermionic raising and lowering operators in terms of anticommuting variables as

$$f^\dagger \rightarrow \xi, \quad f \rightarrow \frac{\partial}{\partial\xi} \quad [24]$$

and the states by

$$|0\rangle \rightarrow 1, \quad |1\rangle \rightarrow \xi \quad [25]$$

Then an arbitrary state takes the form of a linear superposition

$$|\Psi\rangle = \psi_0|0\rangle + \psi_1|1\rangle \rightarrow \Psi(\xi) = \psi_0 + \psi_1\xi \quad [26]$$

and the standard positive-semidefinite inner product on the state space is represented on the wave functions by the double integral

$$\langle \Phi | \Psi \rangle = \int d\xi d\bar{\xi} e^{\bar{\xi}\xi} \Phi^*(\bar{\xi}) \Psi(\xi) = \phi_0^* \psi_0 + \phi_1^* \psi_1 \quad [27]$$

By construction, $f^\dagger = \xi$ and $f = \partial/\partial\xi$ are conjugates with respect to this inner product:

$$\int d\xi d\bar{\xi} e^{\bar{\xi}\xi} \Phi^*(\bar{\xi}) \xi \Psi(\xi) = \int d\bar{\xi} e^{\bar{\xi}\xi} \left(\frac{\partial \Phi}{\partial \xi} \right)^* (\bar{\xi}) \Psi(\xi) \quad [28]$$

The real (self-conjugate) forms of the fermion operators are, therefore, defined by

$$\sigma_1 = \left(\xi + \frac{\partial}{\partial \xi} \right), \quad \sigma_2 = i \left(\xi - \frac{\partial}{\partial \xi} \right) \quad [29]$$

which satisfy the Pauli-Dirac anticommutation relations

$$\sigma_i \sigma_j + \sigma_j \sigma_i = 2\delta_{ij} \quad [30]$$

By taking the product, we obtain

$$\begin{aligned} \sigma_3 &= -i\sigma_1\sigma_2 = 1 - 2\xi \frac{\partial}{\partial \xi} = 1 - 2N_f \\ \Leftrightarrow N_f &= \frac{1}{2}(1 - \sigma_3) \end{aligned} \quad [31]$$

Thus, we may think of the wave functions as two-component spinors, the components being labeled either by the eigenvalues of the spin operator σ_3 , or equivalently by the fermion number N_f , which is a projection operator on the states with negative spin.

The action of the Hamiltonian on a wave function $\Psi(\xi)$ is represented by the integral

$$[H\Psi](\xi) = \int d\xi' d\bar{\xi} e^{\bar{\xi}(\xi' - \xi)} H(\xi, \bar{\xi}) \Psi(\xi') \quad [32]$$

where $H(\xi, \bar{\xi})$ is the ordered symbol of the Hamiltonian:

$$H(\xi, \bar{\xi}) = \varepsilon_f + \hbar\omega\xi\bar{\xi} \quad [33]$$

This expression is to be considered as the classical Hamiltonian of the system. In particular, the exponent of the action

$$\begin{aligned} S &= \int_1^2 dt (i\hbar\bar{\xi}\dot{\xi} - H(\xi, \bar{\xi})) \\ &= \hbar \int_1^2 dt (i\bar{\xi}\dot{\xi} + \omega\xi\bar{\xi}) + \varepsilon_f(t_2 - t_1) \end{aligned} \quad [34]$$

provides the integrand for the path-integral representation of the evolution operator in the quantum theory. The proof is not given here; the reader is referred to the literature. In passing, note that as the anticommuting variables $(\xi, \bar{\xi})$ are taken to be dimensionless, one actually should identify the momentum conjugate to ξ with $\pi = -i\hbar\dot{\xi}$; in the quantum theory, this is replaced by the operator $-i\hbar\partial/\partial\xi$.

Classical Supersymmetry

The classical action for the supersymmetric oscillator with bosonic amplitude x and fermionic amplitude ξ is

$$S = \int_1^2 dt \left(\frac{1}{2} \dot{x}^2 - \frac{\omega}{2} x^2 + i\bar{\xi}\dot{\xi} + \omega\xi\bar{\xi} \right) \quad [35]$$

As inferred from the quantum theory, it is a combination of a linear harmonic oscillator and a fermionic oscillator of the same frequency. A factor $\sqrt{\hbar}$ is also absorbed in ξ and $\bar{\xi}$; equivalently, we can use natural units in which $\hbar = 1$. In the following, we use this convention.

The action [35] is invariant under infinitesimal symmetry transformations

$$\begin{aligned} \delta x &= -i(\bar{\epsilon}\xi + \epsilon\bar{\xi}) \\ \delta\xi &= (\dot{x} + i\omega x)\epsilon, \quad \delta\bar{\xi} = (\dot{x} - i\omega x)\bar{\epsilon} \end{aligned} \quad [36]$$

with $(\bar{\epsilon}, \epsilon)$ Grassmann-odd parameters. The Noether theorem then implies that there are conserved fermionic charges

$$Q = (p - i\omega x)\xi, \quad \bar{Q} = (p + i\omega x)\bar{\xi} \quad [37]$$

with the momentum defined by $p = \dot{x}$. The other conserved quantity is the energy, represented by the Hamiltonian

$$H = \frac{1}{2}(p^2 + \omega^2 x^2) + \omega\xi\bar{\xi} \quad [38]$$

The canonical phase-space formulation is obtained by defining brackets of two functions (A, B) on the phase space $(x, p; \xi, \bar{\xi})$ by

$$\begin{aligned} \{A, B\} &= \frac{\partial A}{\partial x} \frac{\partial B}{\partial p} - \frac{\partial A}{\partial p} \frac{\partial B}{\partial x} \\ &\quad + i(-1)^A \left(\frac{\partial A}{\partial \xi} \frac{\partial B}{\partial \bar{\xi}} + \frac{\partial A}{\partial \bar{\xi}} \frac{\partial B}{\partial \xi} \right) \end{aligned} \quad [39]$$

where $(-1)^A$ is the Grassmann parity of A . In terms of these brackets, the time evolution and supersymmetry transformations take the form

$$\dot{A} = -\{H, A\}, \quad \delta A = i\{\bar{\epsilon}Q + \epsilon\bar{Q}, A\} \quad [40]$$

Moreover, the charges Q and \bar{Q} satisfy the bracket algebra

$$\{Q, \bar{Q}\} = -2iH, \quad \{Q, H\} = \{\bar{Q}, H\} = 0 \quad [41]$$

Thus, the action [35] is the classical counterpart of the quantum theory [9]–[17] in the correspondence limit $i\{A, B\} \rightarrow [A, B]_{\pm} = AB \pm BA$. For these theories, supersymmetry is rooted in the classical transformations [36].

Supersymmetric Quantum Mechanics

The construction for the supersymmetric oscillator can be generalized to other dynamical systems in two ways. First, the nature of the interactions as represented by the potential can be modified. Second, the number of degrees of freedom can be varied. This section presents a generalization of the supersymmetric oscillator to anharmonic interactions, obtained by modification of the supercharges [37] with a general function $\Phi(x)$ as follows:

$$Q = (p - i\Phi(x))\xi, \quad \bar{Q} = (p + i\Phi(x))\bar{\xi} \quad [42]$$

The brackets [39] imply the supersymmetry algebra [41] with the Hamiltonian

$$H = \frac{i}{2}\{Q, \bar{Q}\} = \frac{1}{2}p^2 + \frac{1}{2}\Phi^2(x) + \frac{1}{2}\Phi'(x)(\xi\bar{\xi} - \bar{\xi}\xi) \quad [43]$$

In quantum mechanics, the supercharges become operators Q and Q^\dagger upon reinterpretation of (x, p) as canonically conjugate operators, and the replacement $\xi \rightarrow f^\dagger$ and $\bar{\xi} \rightarrow f$; this procedure involves no ordering ambiguity. The Hamiltonian operator defined by the anticommutator of Q and Q^\dagger then takes the operator form associated with [43]. With the identification

$$A = \frac{1}{\sqrt{2}}(p - i\Phi), \quad A^\dagger = \frac{1}{\sqrt{2}}(p + i\Phi) \quad [44]$$

and making use of the (anti)commutation relations

$$AA^\dagger - A^\dagger A = \Phi'(x), \quad ff^\dagger + f^\dagger f = 1 \quad [45]$$

this Hamilton operator can be written in normal-ordered form as

$$H = \frac{1}{2}(QQ^\dagger + Q^\dagger Q) = A^\dagger A + \Phi'(x)f^\dagger f \quad [46]$$

It is positive-semidefinite by construction. All results for the supersymmetric oscillator are reproduced upon taking $\Phi(x) = \omega x$.

As the Hamiltonian commutes with the fermion number operator N_f , we can label all stationary

states $|E, n_f\rangle$ by the energy E and the fermion number $n_f = (0, 1)$. Moreover, all states of positive energy are degenerate with respect to fermion number, as they form pairs related by supersymmetry:

$$Q|E, 0\rangle = \sqrt{2E}|E, 1\rangle, \quad \bar{Q}|E, 1\rangle = \sqrt{2E}|E, 0\rangle \quad [47]$$

Only ground states with $E_0 = 0$ can occur as singlets under supersymmetry. The existence of such a ground state with fermion number n_f amounts to the existence of a state $|0, n_f\rangle$ satisfying

$$A^\dagger f|0, n_f\rangle = A f^\dagger|0, n_f\rangle = 0 \quad [48]$$

The corresponding wave functions are of the form

$$\begin{aligned} |0, 0\rangle &\rightarrow \Psi_0(x, \xi) = \psi_-(x) \\ |0, 1\rangle &\rightarrow \Psi_1(x, \xi) = \psi_+(x)\xi \end{aligned} \quad [49]$$

where $\psi_{\pm}(x)$ are solutions of the equations

$$A\psi_- = 0, \quad A^\dagger\psi_+ = 0 \quad [50]$$

These functions are formally given by the expressions

$$\psi_{\pm}(x) = C_{\pm} e^{\pm \int_0^x \Phi(y) dy} \quad [51]$$

For a zero-energy ground state to exist, one of these functions must be normalizable. For example, if $\Phi(x)$ is a polynomial of positive odd degree $2k - 1$, then, depending on the sign of the coefficient of x^{2k-1} , one of the exponents is bounded, approaching zero for $x \rightarrow \pm\infty$, and as a result becomes square integrable.

If no normalizable wave functions of the form [51] exist, the ground state cannot have zero energy ($E_0 > 0$) and all states necessarily belong to superdoublets.

Spinning-Particle Mechanics

Minimal supersymmetric classical or quantum mechanics requires equal number of bosonic and fermionic coordinates in configuration space (x_i, ξ_i) , rather than equal number of bosonic and fermionic degrees of freedom in phase space. Specifically, minimal free supersymmetric particle mechanics in n dimensions is described by the classical Lagrangian

$$L = \frac{1}{2}\dot{x}_i^2 + \frac{i}{2}\xi_i\dot{\xi}_i, \quad i = 1, \dots, n \quad [52]$$

It is invariant modulo a total time derivative under infinitesimal supersymmetry transformations

$$\delta x_i = -i\epsilon\xi_i, \quad \delta\xi_i = \dot{x}_i\epsilon \quad [53]$$

The canonical phase-space formulation is phrased in terms of the free-particle momentum and Hamiltonian

$$p_i = \dot{x}_i, \quad H = \frac{1}{2}p_i^2 \quad [54]$$

and the brackets

$$\{A, B\} = \frac{\partial A}{\partial x_i} \frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial B}{\partial x_i} + i(-1)^A \frac{\partial A}{\partial \xi_i} \frac{\partial B}{\partial \xi_i} \quad [55]$$

The supersymmetry transformations are generated by the supercharge

$$Q = p_i \xi_i, \quad \delta A = i\epsilon \{Q, A\} \quad [56]$$

with the supersymmetry algebra

$$i\{Q, Q\} = 2H, \quad \{Q, H\} = 0 \quad [57]$$

An important quantity in these models is the bilinear (Grassmann-even) antisymmetric tensor

$$\sigma_{ij} = i\xi_i \xi_j \quad [58]$$

For a free particle, it is a set of constants of motion forming a representation of $so(n)$, the Lie algebra of n -dimensional rotations:

$$\{\sigma_{ij}, \sigma_{kl}\} = \delta_{jk}\sigma_{il} - \delta_{jl}\sigma_{ik} - \delta_{ik}\sigma_{jl} + \delta_{il}\sigma_{jk} \quad [59]$$

Therefore, the physical interpretation of σ_{ij} is that it represents the particle spin. For this reason, supersymmetric particle mechanics is often called spinning-particle mechanics.

Quantum mechanics of the spinning particle has the same algebraic structure, with (x_i, p_i) the standard canonically conjugate operators, and the fermionic coordinates ξ_i represented by the generators of a Clifford algebra; the irreducible representation in terms of Pauli-Dirac matrices of dimension $2^{[n/2]} \times 2^{[n/2]}$ is

$$\xi_i \rightarrow \frac{1}{\sqrt{2}}\gamma_i, \quad \gamma_i \gamma_j + \gamma_j \gamma_i = 2\delta_{ij} \quad [60]$$

It follows that the wave functions have $2^{[n/2]}$ components, describing different polarization states. Furthermore, in minimal supersymmetric quantum mechanics, the supersymmetry operator is represented by the Dirac operator:

$$Q \rightarrow \frac{1}{\sqrt{2}}\gamma \cdot p, \quad (\gamma \cdot p)^2 = p_i^2 = 2H \quad [61]$$

Hence, the stationary states of the system solve the Dirac equation

$$\gamma \cdot p \Psi = \sqrt{2E} \Psi \quad [62]$$

The models can, without difficulty, be extended to include interactions with external fields. As an example, we consider the coupling to a magnetic

field described by a vector potential $A_i(x)$. An extension of the free-particle action [52], invariant under the same supersymmetry transformations [53], is

$$S = \int dt \left(\frac{1}{2} \dot{x}_i^2 + \frac{i}{2} \xi_i \dot{\xi}_i + q A_i(x) \dot{x}_i - \frac{iq}{2} F_{ij}(x) \xi_i \xi_j \right) \quad [63]$$

where $F_{ij} = \nabla_i A_j - \nabla_j A_i$ is the field strength. The canonical momentum in this model is

$$p_i = \dot{x}_i + q A_i(x) \quad [64]$$

with the result that the canonical expressions for the Hamiltonian and supercharge become

$$H = \frac{1}{2}(p_i - q A_i(x))^2, \quad Q = (p_i - q A_i(x)) \xi_i \quad [65]$$

In the quantum theory, these constants of motion become the covariant Laplacian and Dirac operator in an external vector potential $A_i(x)$. Observe that supersymmetry requires the spin to couple to the magnetic field with gyromagnetic ratio $g=2$. Explicitly, the equation of motion for ξ can be transformed into an equation for the spin precession:

$$\dot{\xi}_i = q F_{ij} \xi_j \Rightarrow \dot{\sigma}_{ij} = q(F_{ik} \sigma_{kj} - \sigma_{ik} F_{kj}) \quad [66]$$

In three dimensions, this is equivalent to an equation in terms of axial vectors:

$$F_{ij} = \varepsilon_{ijk} B_k, \quad \sigma_{ij} = \varepsilon_{ijk} s_k \Rightarrow \dot{s} = -q B \times s \quad [67]$$

showing that the precession rate of s is given by twice the Larmor frequency.

Extended Supersymmetry

It is possible to construct theories with more supersymmetries by associating with every bosonic coordinate several fermionic coordinates. An example is the supersymmetric oscillator and its generalizations considered earlier, which has equal number of bosonic and fermionic degrees of freedom in phase space, rather than equal number of bosonic and fermionic coordinates in configuration space. The classical phase space, spanned by variables $(x_i, p_i; \xi_i, \bar{\xi}_i)$ with $i=1, \dots, n$, then has double the number of fermionic variables compared to the minimal supersymmetric particle models. Such models can be constructed for systems with an n -dimensional bosonic configuration space. Their supercharges take the form

$$Q = (p_i - i\Phi_i(x)) \xi_i, \quad \bar{Q} = (p_i + i\Phi_i(x)) \bar{\xi}_i \quad [68]$$

$$x = (x_1, \dots, x_n)$$

whilst the Hamiltonian becomes

$$H = \frac{1}{2}p_i^2 + \frac{1}{2}\Phi_i^2(\mathbf{x}) + \frac{1}{4}(\nabla_i\Phi_i + \nabla_i\Phi_i)(\xi_i\bar{\xi}_i - \bar{\xi}_i\xi_i) \quad [69]$$

The supercharges are conserved if the curl of $\Phi_i(\mathbf{x})$ vanishes: $\nabla_i\Phi_j - \nabla_j\Phi_i = 0$. It follows that at least locally there exists a single function $W(\mathbf{x})$ such that

$$\Phi_i(\mathbf{x}) = \nabla_i W(\mathbf{x}) \quad [70]$$

$W(\mathbf{x})$ is called the superpotential. Defining the operators

$$A_i = p_i - i\Phi_i(\mathbf{x}), \quad A_i^\dagger = p_i + i\Phi_i(\mathbf{x}) \quad [71]$$

$$A_i A_j^\dagger - A_j^\dagger A_i = \nabla_i\Phi_j + \nabla_j\Phi_i$$

the supersymmetric quantum theory is defined by

$$Q = A_i f_i^\dagger, \quad Q^\dagger = A_i^\dagger f_i \quad [72]$$

$$H = \frac{1}{2}(QQ^\dagger + Q^\dagger Q)$$

The Hamiltonian is the direct operator translation of the classical expression [69]; its normal-ordered form is

$$H = A_i^\dagger A_i + \frac{1}{2}(\nabla_i\Phi_j + \nabla_j\Phi_i)f_i^\dagger f_j \quad [73]$$

The total fermion number operator

$$N_f = f_i^\dagger f_i \quad [74]$$

(summed over i) satisfying the commutation relations

$$[N_f, f_i^\dagger]_- = f_i^\dagger, \quad [N_f, f_i]_- = -f_i \quad [75]$$

commutes with the Hamiltonian. Hence, the stationary states can be labeled by the energy E and the total fermion number $n_f = (0, \dots, n)$. The energy spectrum being positive semidefinite, all positive-energy states occur in pairs of fermion number $(n_f, n_f + 1)$; zero-energy states exist only if the equations

$$A_i f_i^\dagger |0, n_f\rangle = A_i^\dagger f_i |0, n_f\rangle = 0 \quad [76]$$

admit a normalizable solution. In this context, the vanishing of the curl of $\Phi_i(\mathbf{x})$ is important, as it is a necessary condition for the formal solutions

$$\psi_\pm(\mathbf{x}) = C_\pm \exp\left(\pm \int_0^{\mathbf{x}} \Phi(\mathbf{y}) \cdot d\mathbf{y}\right) = C'_\pm e^{\pm W(\mathbf{x})} \quad [77]$$

to be single-valued. If one of them is normalizable, there exists a zero-energy ground state with $n_f = 0$ or $n_f = n$, represented by a wave function:

$$|0, 0\rangle \rightarrow \Psi_0(\mathbf{x}, \xi) = \psi_-(\mathbf{x}) \quad [78]$$

$$|0, n\rangle \rightarrow \Psi_n(\mathbf{x}, \xi) = \psi_+(\mathbf{x}) \xi_1 \dots \xi_n$$

Alternatively, we can represent the wave functions as spinors of dimension 2^n , on which the fermion operators f_i^\dagger and f_i act as a 2^n -dimensional matrix representation of the Clifford algebra with generators $\gamma_a, a = 1, \dots, 2n$, defined by

$$\gamma_i = f_i + f_i^\dagger, \quad \gamma_{i+r} = i(f_i - f_i^\dagger) \quad [79]$$

These operators indeed satisfy the anticommutation rule

$$\gamma_a \gamma_b + \gamma_b \gamma_a = 2\delta_{ab} \quad [80]$$

Thus, the wave functions have 2^n components, as compared to the $2^{[n/2]}$ polarization states of the minimal models.

The Witten Index

We have noted that for supersymmetric quantum systems, like the harmonic and anharmonic supersymmetric oscillator, states exist in pairs of different fermion number, degenerate in energy, except for possibly one or more zero-energy states which are superinvariant in the sense that

$$Q|0, n\rangle = \bar{Q}|0, n\rangle = 0 \Leftrightarrow H|0, n\rangle = 0 \quad [81]$$

In the Schrödinger representation, these states are characterized as zero modes of the Dirac operator:

$$\gamma \cdot D \Psi = 0 \quad [82]$$

where D_i is an ordinary or field-dependent (e.g., covariant) derivative. Clearly, the existence of such states can, in some cases, be guaranteed if there is no state which can pair up with a given state to form a superdoublet. Witten developed a topological characterization of this condition, encoded in an index defined by

$$I = \text{tr}(-1)^{N_f} = n_b(E=0) - n_f(E=0) \quad [83]$$

where N_f is the fermion number operator, and $n_{b,f}(E=0)$ are the number of bosonic and fermionic zero-energy states. The trace is taken over the complete space of states, but as all nonzero energy states occur in pairs of a bosonic and a fermionic state, their contributions to the trace cancel, having opposite sign. Therefore, the trace is actually only over the zero-energy states, and counts the number of bosonic states with positive sign, and the number of fermionic states with negative sign. If the index vanishes, $I=0$, then any zero-energy states necessarily exist in equal number of bosonic and fermionic states; under perturbations of the potential, these states can form pairs and change their energy to a positive value. However, if the index does not vanish, $I \neq 0$, then there are states which have no

partner of complementary fermion number; these states can never get a nonzero energy under changes in the parameters of the potential, as long as the changes respect supersymmetry. Such systems, therefore, necessarily possess exact zero-energy states which are invariant under all supersymmetries.

Deformations of the potential respecting supersymmetry are those obtained by changing the parameters in the superpotential. The usefulness of this concept is, therefore, that the index for models with complicated superpotentials can be computed by comparing them with models with simple superpotentials having similar topological properties.

Counting the number of states is not always a simple procedure, in particular when the spectrum includes continuum states. Therefore, in practice one often needs a regularization procedure, by taking the trace over the full state space of the exponentially damped quantity

$$I(\beta) = \text{tr}(-1)^{N_f} e^{-\beta H} \quad [84]$$

and taking the limit $\beta \rightarrow 0$. The quantity [84] can be computed in terms of a path integral with periodic boundary conditions for the fermionic degrees of freedom.

Finally, as the wave function representation of supersymmetric quantum mechanics [82] links the Witten index to the space of zero modes of a Dirac operator, in particular cases it can be used to describe topological aspects of sigma models and gauge theories, and related mathematical quantities such as the Atiyah–Singer index.

More details and references to the original literature can be found in the reviews listed in the Further Reading section.

See also: Path-Integrals in Non Commutative Geometry; Supermanifolds.

Further Reading

- Cooper F, Khare A, and Sukhatme U (1995) Supersymmetry and quantum mechanics. *Physics Reports* 251: 267.
 De Witt BS (1984, 1992) *Supermanifolds*. Cambridge: Cambridge University Press.
 Shifman MA (1999) *ITEP Lectures on Particle Physics and Field Theory*, vol. 1, ch. 4. Singapore: World Scientific.
 van Holten JW (1996) D=1 Supergravity and spinning particles. In: Jancewicz B and Sobezyk J (eds.) *From Field Theory to Quantum Groups*, p. 173. Singapore: World Scientific.

Supersymmetry Methods in Random Matrix Theory

M R Zirnbauer, Universität Köln, Köln, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

A prominent theme of modern condensed matter physics is electronic transport – in particular, the electrical conductivity – of disordered metallic systems at very low temperatures. From the Landau theory of weakly interacting Fermi liquids, one expects the essential aspects of the situation to be captured by the single-electron approximation. Mathematical models that have been proposed and studied in this context include random Schrödinger operators and band random matrices.

If the physical system has infinite size, two distinct possibilities exist: the quantum single-electron motion may either be bounded or unbounded. In the former case, the disordered electron system is an insulator, in the latter case, a metal with finite conductivity (if the electron motion is not critical but diffusive). Metallic behavior is expected for weakly disordered systems in three dimensions;

insulating behavior sets in when the disorder strength is increased or the space dimension reduced.

The main theoretical tool used in the physics literature on the subject is the “supersymmetry method” pioneered by Wegner and Efetov (1979–83). Over the past 20 years, physicists have applied the method in many instances, and a rather complete picture of weakly disordered metals has emerged. Several excellent reviews of these developments are available in print.

From the perspective of mathematics, however, the method has not always been described correctly, and what is sorely lacking at present is an exposition of how to implement the method rigorously. (Unfortunately, the correct exposition by Schäfer and Wegner (1980) was largely ignored or forgotten by later authors.) In this article, an attempt is made to help remedy the situation, by giving a careful review of the Wegner–Efetov supersymmetry method for the case of Hermitian band random matrices.

Gaussian Ensembles

Let V be a unitary vector space of finite dimension. A Hermitian random matrix model on V is defined

by some probability distribution on $\text{Herm}(V)$, the Hermitian linear operators on V . We may fix some orthonormal basis of V and represent the elements H of $\text{Herm}(V)$ by Hermitian square matrices.

Quite generally, probability distributions are characterized by their Fourier transform or characteristic function. In the present case this is

$$\Omega(K) = \langle e^{i\text{tr}HK} \rangle$$

where the Fourier variable K is some other linear operator on V , and $\langle \dots \rangle$ denotes the expectation value with respect to the probability distribution for H . Later, it will be important that, if $\Omega(K)$ is an analytic function of K , the matrix entries of K need not be from \mathbb{R} or \mathbb{C} but can be taken from the even part of some exterior algebra.

The probability distributions to be considered in this article are Gaussian with zero mean, $\langle H \rangle = 0$. Their Fourier transform is also Gaussian:

$$\Omega(K) = e^{-(1/2)J(K,K)}$$

with J some quadratic form. We now describe J for a large family of hierarchical models that includes the case of band random matrices.

Let V be given a decomposition by orthogonal vector spaces:

$$V = V_1 \oplus V_2 \oplus \dots \oplus V_{|\Lambda|}$$

We should imagine that every vector space V_i corresponds to one site i of some lattice Λ , and the total number of sites is $|\Lambda|$. For simplicity, we take all dimensions to be equal: $\dim V_1 = \dots = \dim V_{|\Lambda|} = N$. Thus, the dimension of V is $N|\Lambda|$. The integer N is called the number of orbitals per site.

If Π_i is the orthogonal projector on the linear subspace $V_i \subset V$, we take the bilinear form J to be

$$J(K, K') = \sum_{i,j=1}^{|\Lambda|} J_{ij} \text{tr}(\Pi_i K \Pi_j K')$$

where the coefficients J_{ij} are real, symmetric, and positive. This choice of J implies invariance under the group \mathcal{U} of unitary transformations in each subspace:

$$\mathcal{U} = \text{U}(V_1) \times \text{U}(V_2) \times \dots \times \text{U}(V_{|\Lambda|})$$

Clearly, $\Omega(K) = \Omega(UKU^{-1})$ or, equivalently, the probability distribution for H is invariant under conjugation $H \mapsto UHU^{-1}$, for $U \in \mathcal{U}$.

If $\{e_i^a\}_{a=1,\dots,N}$ is an orthonormal basis of V_i , we define linear operators $E_{ij}^{ab} : V_j \rightarrow V_i$ by $E_{ij}^{ab} e_j^b = e_i^a$. By evaluating $J(E_{ij}^{ab}, E_{j'i'}^{b'a'}) = J_{ij} \delta_{ii'} \delta_{jj'} \delta^{aa'} \delta^{bb'}$, one sees

that the matrix entries of H all are statistically independent.

By varying the lattice Λ , the number of orbitals N , and the variances J_{ij} , one obtains a large class of Hermitian random matrix models, two prominent subclasses of which are the following:

1. For $|\Lambda|=1$, one gets the Gaussian Unitary Ensemble (GUE). Its symmetry group is $\mathcal{U} = \text{U}(N)$, the largest one possible in dimension $N = \dim V$.
2. If $|i-j|$ denotes a distance function for Λ , and f a rapidly decreasing positive function on \mathbb{R}_+ of width W , the choice $J_{ij} = f(|i-j|)$ with $N=1$ gives an ensemble of band random matrices with bandwidth W and symmetry group $\mathcal{U} = \text{U}(1)^{|\Lambda|}$.

Beyond being real, symmetric, and positive, the variances J_{ij} are required to have two extra properties in order for all of the following treatment to go through:

- They must be positive as a quadratic form. This is to guarantee the existence of an inverse, which we denote by $w_{ij} = (J^{-1})_{ij}$.
- The off-diagonal matrix entries of the inverse must be nonpositive: $w_{ij} \leq 0$ for $i \neq j$.

Basic Tools

Green's Functions

A major goal of random matrix theory is to understand the statistical behavior of the spectrum and the eigenstates of a random Hamiltonian H . Spectral and eigenstate information can be extracted from the Green's function, that is, from matrix elements of the operator $(z - H)^{-1}$ with complex parameter $z \in \mathbb{C} \setminus \mathbb{R}$. For the models at hand, the good objects to consider are averages of \mathcal{U} -invariant observables such as

$$G_i^{(1)}(z) = \langle \text{tr} \Pi_i (z - H)^{-1} \rangle \quad [1]$$

$$G_{ij}^{(2)}(z_1, z_2) = \langle \text{tr} \Pi_i (z_1 - H)^{-1} \Pi_j (z_2 - H)^{-1} \rangle \quad [2]$$

The discontinuity of $G_i^{(1)}(z)$ across the real z -axis yields the local density of states. In the limit of infinite volume ($|\Lambda| \rightarrow \infty$), the function $G_{ij}^{(2)}(z_1, z_2)$ for $z_1 = E + i\varepsilon$, $z_2 = E - i\varepsilon$, real energy E , and $\varepsilon > 0$ going to zero, gives information on transport, for example, the electrical conductivity by the Kubo–Greenwood formula.

Mathematically speaking, if $G_{ij}^{(2)}(E + i\varepsilon, E - i\varepsilon)$ is bounded (for infinite volume) in ε and decreases algebraically with distance $|i-j|$ at $\varepsilon=0+$, the

spectrum is absolutely continuous and the eigenstates are extended at energy E . On the other hand, a pure point spectrum and localized eigenstates are signaled by the behavior $G_{ij}^{(2)} \sim \varepsilon^{-1} e^{-\lambda|i-j|}$ with positive Lyapunov exponent λ .

Green's Functions from Determinants

For any pair of linear operators A, B on a finite-dimensional vector space V , the following formula from basic linear algebra holds if A has an inverse:

$$\left. \frac{d}{dt} \det(A + tB) \right|_{t=0} = \det(A) \operatorname{tr}(A^{-1}B)$$

Using it with $A = z - H$ and $z \in \mathbb{C} \setminus \mathbb{R}$, all Green's functions can be expressed in terms of determinants; for example,

$$\begin{aligned} G_{ij}^{(2)}(w, z) &= \sum_{a,b=1}^N \frac{\partial^2}{\partial s \partial t} \left\langle \frac{\det(w - H) \det(z - H + tE_{ij}^{ab})}{\det(w - H - sE_{ji}^{ba}) \det(z - H)} \right\rangle \Big|_{s=t=0} \end{aligned}$$

It is clear that, given a formula of this kind, what one wants is a method to handle ensemble averages of ratios of determinants. This is what is reviewed in the sequel.

Determinants as Gaussian Integrals

Let the Hermitian scalar product of the unitary vector space V be written as $\varphi_1, \varphi_2 \mapsto (\bar{\varphi}_1, \varphi_2)$, and denote the adjoint or Hermitian conjugate of a linear operator A on V by A^* . If $\Re A := (1/2)(A + A^*) > 0$, the standard Lebesgue integral of the Gaussian function $\varphi \mapsto e^{-(\bar{\varphi}, A\varphi)}$ makes sense and gives

$$\int e^{-(\bar{\varphi}, A\varphi)} = \det A^{-1} \quad [3]$$

where it is understood that we are integrating with the Lebesgue measure on (the normed vector space) V normalized by $\int e^{-(\bar{\varphi}, \varphi)} = 1$. The same integral with anticommuting ψ instead of the (commuting) $\varphi \in V$ gives

$$\int e^{-(\bar{\psi}, A\psi)} = \det A \quad [4]$$

This basic formula from the field theory of fermionic particles is a consequence of the integration over anticommuting variables actually being *differentiation*:

$$\int d\bar{\psi}_1 d\psi_1 f(\bar{\psi}_1, \psi_1, \dots) := \frac{\partial^2}{\partial \bar{\psi}_1 \partial \psi_1} f(\bar{\psi}_1, \psi_1, \dots)$$

Fermionic Variant

The supersymmetry method of random matrix theory is a theme with many variations. The first variation to be described is the “fermionic” one. To optimize the notation, we now write $d\mu_{N,J}(H)$ for the density of the Gaussian probability distribution of H :

$$\langle F(H) \rangle = \int F(H) d\mu_{N,J}(H)$$

All determinants and traces appearing below will be taken over vector spaces that are clear from the context.

Let z_1, \dots, z_n be any set of n complex numbers, put $z := \operatorname{diag}(z_1, \dots, z_n)$ for later purposes, and consider

$$\Omega_{n,N}^{\text{ferm}}(z, J) = \int \prod_{\alpha=1}^n \det(z_\alpha - H) d\mu_{N,J}(H) \quad [5]$$

The supersymmetry method expresses this average of a product of determinants in an alternative way, by integrating over a “dual” measure as follows.

Introducing an auxiliary unitary vector space \mathbb{C}^n , one associates with every site i of the lattice Λ an object $Q_i \in \operatorname{Herm}(\mathbb{C}^n)$, the space of Hermitian $n \times n$ matrices. If dQ_i for $i = 1, \dots, |\Lambda|$ are Lebesgue measures on $\operatorname{Herm}(\mathbb{C}^n)$, one puts $DQ = \text{const.} \times \prod_i dQ_i$ and

$$d\nu_{n,J}(Q) := e^{-(1/2)\Sigma_{ij}(J^{-1})_{ij} \operatorname{tr} Q_i Q_j} DQ \quad [6]$$

The multiplicative constant in DQ is fixed by requiring the density to be normalized: $\int d\nu_{n,J}(Q) = 1$. By completing the square, this Gaussian probability measure has the characteristic function

$$\int e^{i\Sigma_j \operatorname{tr} Q_j K_j} d\nu_{n,J}(Q) = e^{-(1/2)\Sigma_{ij} J_{ij} \operatorname{tr} K_i K_j}$$

where the Fourier variables $K_1, \dots, K_{|\Lambda|}$ are $n \times n$ matrices with matrix entries taken from \mathbb{C} or another commutative algebra.

The key relation of the fermionic variant of the supersymmetry method is that the expectation of the product of determinants [5] has another expression as

$$\Omega_{n,N}^{\text{ferm}}(z, J) = \int \prod_{j=1}^{|\Lambda|} \det^N(z - iQ_j) d\nu_{n,J}(Q) \quad [7]$$

($i = \sqrt{-1}$). The strategy of the proof is quite simple: one writes the determinants in both expressions for $\Omega_{n,N}^{\text{ferm}}$ as Gaussian integrals over $nN|\Lambda|$ complex fermionic variables ψ_1, \dots, ψ_n (each ψ_α is a vector in V with anticommuting coefficients), using the basic

formula [4]. The integrals then encountered are essentially the Fourier transforms of the distributions $d\mu_{N,J}(H)$ resp., $d\nu_{n,J}(Q)$. The result is

$$\int e^{-\sum_{\gamma} z_{\gamma}(\bar{\psi}_{\gamma}, \psi_{\gamma})} e^{-(1/2)\sum_{ij} J_{ij} \sum_{\alpha\beta} (\bar{\psi}_{\alpha}, \Pi_i \psi_{\beta})(\bar{\psi}_{\beta}, \Pi_j \psi_{\alpha})}$$

for both expressions of $\Omega_{n,N}^{\text{ferm}}$. In other words, although the probability distributions $d\mu_{N,J}(H)$ and $d\nu_{n,J}(Q)$ are distinct (they are defined on different spaces), their characteristic functions coincide when evaluated on the Fourier variables $K = \sum_{\alpha} \psi_{\alpha}(\bar{\psi}_{\alpha}, \bullet)$ for H and $(K_i)_{\alpha\beta} = (\bar{\psi}_{\alpha}, \Pi_i \psi_{\beta})$ for Q_i . This establishes the claimed equality of the expressions [5] and [7] for $\Omega_{n,N}^{\text{ferm}}(z, J)$.

What is the advantage of passing to the alternative expression by $d\nu_{n,J}(Q)$? The answer is that, while H is made up of independent random variables, the new variables Q_i , called the Hubbard–Stratonovich field, are correlated: they interact through the “exchange” constants $w_{ij} = (J^{-1})_{ij}$. If that interaction creates enough collectivity, a kind of mean-field behavior results.

For the simple case of GUE ($|\Lambda| = 1, w_{11} = N/\lambda^2$) with $z_1 = \dots = z_n = E$, one gets the relation

$$\langle \det^n(E - H) \rangle = \int \det^N(E - iQ) e^{-(N/2\lambda^2)\text{tr } Q^2} dQ$$

the right-hand side of which is easily analyzed by the steepest descent method in the limit of large N .

For band random matrices in the so-called ergodic regime, the physical behavior turns out to be governed by the constant mode $Q_1 = \dots = Q_{|\Lambda|}$ – a fact that can be used to establish GUE universality in that regime.

Bosonic Variant

The bosonic variant of the present method, due to Wegner, computes averages of products of determinants placed in the denominator:

$$\Omega_{n,N}^{\text{bos}}(z, J) = \int \prod_{\alpha=1}^n \det^{-1}(z_{\alpha} - H) d\mu_{N,J}(H) \quad [8]$$

where we now require $\Im m z_{\alpha} \neq 0$ for all $\alpha = 1, \dots, n$. Complications relative to the fermionic case arise from the fact that the integrand in [8] has poles. If one replaces the anticommuting vectors ψ_{α} by commuting ones φ_{α} , and then simply repeats the previous calculation in a naive manner, one arrives at

$$\Omega_{n,N}^{\text{bos}}(z, J) \stackrel{?}{=} \int \prod_{j=1}^{|\Lambda|} \det^{-N}(z - Q_j) d\nu_{n,J}(Q) \quad [9]$$

where the integral is still over $Q_j \in \text{Herm}(\mathbb{C}^n)$. The calculation is correct, and relation [9] therefore

holds true, provided that the parameters z_1, \dots, z_n all lie in the same half (upper or lower) of the complex plane. To obtain information on transport properties, however, one needs parameters in both the upper and lower halves; see the paragraph following [2]. The general case to be addressed below is $\Im m z_{\alpha} > 0$ for $\alpha = 1, \dots, p$, and $\Im m z_{\alpha} < 0$ for $\alpha = p + 1, \dots, n$. Careful inspection of the steps leading to eqn [9] reveals a convergence problem for $0 < p < n$. In fact, [9] with Q_j in $\text{Herm}(\mathbb{C}^n)$ turns out to be false in that range. Learning how to resolve this problem is the main step toward mathematical mastery of the method. Let us therefore give the details.

If $s_{\alpha} := \text{sgn} \Im m z_{\alpha}$, the good (meaning convergent) Gaussian integral to consider is

$$\int e^{i \sum_{\alpha} s_{\alpha} (\varphi_{\alpha}, (z_{\alpha} - H) \varphi_{\alpha})} = \prod_{\alpha=1}^n \det^{-1}(-i s_{\alpha} (z_{\alpha} - H))$$

To avoid carrying around trivial constants, we now assume $i^{(n-2p)N/|\Lambda|} = 1$. Use of the characteristic function of the distribution for H then gives

$$\Omega_{n,N}^{\text{bos}}(z, J) = \int e^{i \sum_{\gamma} s_{\gamma} z_{\gamma} (\bar{\varphi}_{\gamma}, \varphi_{\gamma})} \times e^{-\frac{1}{2} \sum_{ij} J_{ij} \sum_{\alpha\beta} s_{\alpha} (\bar{\varphi}_{\alpha}, \Pi_i \varphi_{\beta}) s_{\beta} (\bar{\varphi}_{\beta}, \Pi_j \varphi_{\alpha})} \quad [10]$$

The difficulty of analyzing this expression stems from the “hyperbolic” nature (due to the indefiniteness of the signs $s_{\alpha} = \pm 1$) of the term quartic in the $\varphi_{\alpha}, \bar{\varphi}_{\alpha}$.

Fyodorov’s Method

The integrand for Ω^{bos} is naturally expressed in terms of $n \times n$ matrices M_i with matrix elements $(M_i)_{\alpha\beta} = (\bar{\varphi}_{\alpha}, \Pi_i \varphi_{\beta})$. These matrices lie in $\text{Herm}^+(\mathbb{C}^n)$, that is, they are non-negative as well as Hermitian. Fyodorov’s idea was to introduce them as the new variables of integration. To do that step, recall the basic fact that, given two differentiable spaces X and Y and a smooth map $\psi: X \rightarrow Y$, a distribution μ on X is pushed forward to a distribution $\psi(\mu)$ on Y by $\psi(\mu)[f] := \mu[f \circ \psi]$, where f is any test function on Y .

We apply this universal principle to the case at hand by identifying X with V^n , and Y with $(\text{Herm}^+(\mathbb{C}^n))^{| \Lambda |}$, and ψ with the mapping that sends

$$(\varphi_1, \dots, \varphi_n) \in X \quad \text{to} \quad (M_1, \dots, M_{|\Lambda|}) \in Y$$

by $(M_i)_{\alpha\beta} = (\bar{\varphi}_{\alpha}, \Pi_i \varphi_{\beta})$. On $X = V^n$ we are integrating with the product Lebesgue measure normalized by $\int e^{-\sum_{\alpha} (\bar{\varphi}_{\alpha}, \varphi_{\alpha})} = 1$. We now want the push-forward of this flat measure (or distribution) by the mapping ψ . In general, the push-forward of a measure is not

guaranteed to have a density but may be singular (like a Dirac δ -distribution). This is in fact what happens if $N < n$. The matrices M_i then have less than the maximal rank, so they fail to be positive but possess zero eigenvalues, which implies that the flat measure on X is pushed forward by ψ into the boundary of Y . For $N \geq n$, on the other hand, the push-forward measure does have a density on Y ; and that density is $\prod_{i=1}^{|\Lambda|} (\det M_i)^{N-n} dM_i$, as is seen by transforming to the eigenvalue representation and comparing Jacobians. The dM_i are Lebesgue measures on $\text{Herm}(\mathbb{C}^n)$, normalized by the condition

$$\int_{M_i > 0} e^{-\text{tr} M_i} (\det M_i)^{N-n} dM_i = \int e^{-\sum_{\alpha} (\bar{\varphi}_{\alpha}, \Pi_i \varphi_{\alpha})} = 1$$

Assembling the sign information for $\Im z_{\alpha}$ in a diagonal matrix $s := \text{diag}(s_1, \dots, s_n)$, and pushing the integral over X forward to an integral over Y with measure $DM := \prod_i dM_i$, we obtain Fyodorov's formula:

$$\Omega_{n,N}^{\text{bos}}(z, J) = \int_Y e^{-(1/2) \sum_{ij} J_{ij} \text{tr}(s M_i s M_j)} \times e^{\sum_k \text{tr}(i s z M_k + (N-n) \ln M_k)} DM \quad [11]$$

This formula has a number of attractive features. One is ease of derivation, another is ready generalizability to the case of non-Gaussian distributions. The main disadvantage of the formula is that it does not apply to the case of band random matrices (because of the restriction $N \geq n$); nor does it combine nicely with the fermionic formula [7] to give a supersymmetric formalism, as one formula is built on J_{ij} and the other on w_{ij} .

Note that [11] clearly displays the dependence on the signature of $\Im z$: you cannot remove the s_1, \dots, s_n from the integrand without changing the domain of integration $Y = (\text{Herm}^+(\mathbb{C}^n))^{|\Lambda|}$. This important feature is missing from the naive formula [9].

Setting $q = n - p$, let $U(p, q)$ be the pseudounitary group of complex $n \times n$ matrices T with inverse $T^{-1} = s T^* s$. Since $|\det T| = 1$ for $T \in U(p, q)$, the integration domain Y and density $DM = \prod_i dM_i$ of Fyodorov's formula are invariant under $U(p, q)$ transformations $M_i \mapsto T M_i T^*$, and so is actually the integrand in the limit where all parameters z_1, \dots, z_n become equal. Thus, the elements of $U(p, q)$ are global symmetries in that limit. This observation holds the key to another method of transforming the expression [10].

The Method of Schäfer and Wegner

To rescue the naive formula [9], what needs to be abandoned is the integration domain $\text{Herm}(\mathbb{C}^n)$ for the matrices Q_i . The good domain to use was

constructed by Schäfer and Wegner, but was largely forgotten in later physics work.

Writing $(M_k)_{\alpha\beta} = (\bar{\varphi}_{\alpha}, \Pi_k \varphi_{\beta})$ as before, consider the function

$$F_M(Q) = e^{(1/2) \sum_{ij} w_{ij} \text{tr}(s Q_i + i z)(s Q_j + i z) - \sum_k \text{tr} M_k Q_k} \quad [12]$$

viewed as a holomorphic function of

$$Q = (Q_1, \dots, Q_{|\Lambda|}) \in \text{End}(\mathbb{C}^n)^{|\Lambda|}$$

If the Gaussian integral $\int F_M(Q) DQ$ with holomorphic density $DQ = \prod_i dQ_i$ is formally carried out by completing the square, one gets the integrand of [10]. This is just what we want, as it would allow us to pass to a Q -matrix formulation akin to the one of the previous section. But how can that formal step be made rigorous? To that end, one needs to (1) construct a domain on which $|F_M(Q)|$ decreases rapidly so that the integral exists, and (2) justify completion of the square and shifting of variables.

To begin, take the absolute value of $F_M(Q)$. Putting $(1/2)(Q_j + Q_j^*) =: \Re Q_j$ and $(1/2i)(Q_j - Q_j^*) =: \Im Q_j$, we have $|F_M| = e^{-(1/4)(f_1 + f_2 + f_3)}$ with

$$f_1(Q) = \sum_{ij} w_{ij} \text{tr}(s \Im Q_i + z)(s \Im Q_j + z) + \text{c.c.}$$

$$f_2(Q) = -2 \sum_{ij} w_{ij} \text{tr}(s \Re Q_i)(s \Re Q_j)$$

$$f_3(Q) = 4 \sum_i \text{tr} \left(M_i + s \Im z \sum_j w_{ij} \right) \Re Q_i$$

These expressions suggest making the following choice of integration domain for $Q_i (i = 1, \dots, |\Lambda|)$. Pick some real constant $\lambda > 0$ and put

$$\Re Q_i = \lambda T_i T_i^*, \quad \Im Q_i = P_i := \begin{pmatrix} P_i^+ & 0 \\ 0 & P_i^- \end{pmatrix}$$

with $T_i \in U(p, q)$, $P_i^+ \in \text{Herm}(\mathbb{C}^p)$, $P_i^- \in \text{Herm}(\mathbb{C}^q)$. The set of matrices Q_i so defined is referred to as the Schäfer–Wegner domain $X_{\lambda}^{p,q}$. The range of the field $Q = (Q_1, \dots, Q_{|\Lambda|})$ is the direct product $\mathcal{Z} := (X_{\lambda}^{p,q})^{|\Lambda|}$.

To show that this is a good choice of domain, we first of all show convergence of the integral $\int_{\mathcal{Z}} F_M(Q) DQ$. The matrices P_i commute with s , so

$$f_1(Q)|_{\mathcal{Z}} = 2 \Re \sum_{ij} w_{ij} \text{tr}(P_i + s z)(P_j + s z)$$

Since the coefficients w_{ij} are positive as a quadratic form, this expression is convex (with a positive Hessian) in the Hermitian matrices P_i . Second, the function

$$f_2(Q)|_{\mathcal{Z}} = -2 \lambda^2 \sum_{ij} w_{ij} \text{tr}(T_i T_i^*)^{-1} T_j T_j^*$$

is bounded from below by the constant $-2\lambda^2 n \sum_i w_{ii}$. This holds true because w_{ij} is negative for $i \neq j$, and because $T_i T_i^* > 0$ and the trace of a product of two positive Hermitian matrices is always positive. Third,

$$f_3(Q)|_{\mathcal{J}} = 4\lambda \sum_i \text{tr} \left(M_i + s \mathcal{I} m z \sum_j w_{ij} \right) T_i T_i^*$$

is positive, as (\dots) is positive Hermitian. As long as $s \mathcal{I} m z > 0$, the function f_3 goes to infinity for all possible directions of taking the T_i to infinity on $U(p, q)$.

Thus, when the matrices Q_i are taken to vary on the Schäfer–Wegner domain $X_{\lambda}^{p,q}$, the absolute value $|F_M| = e^{-(1/4)(f_1+f_2+f_3)}$ decreases rapidly at infinity. This establishes the convergence of $\int_{\mathcal{J}} F_M(Q) DQ$.

Next, let us count dimensions. The mapping $T \mapsto TT^*$ for $T \in U(p, q) =: G$ is invariant under right multiplication of T by elements of the unitary subgroup $H := U(p) \times U(q)$ – it is called the “Cartan embedding” of G/H into G . The real manifold G/H has dimension $2pq$ and so does its image under the Cartan embedding. Augmenting this by the dimension of $\text{Herm}(\mathbb{C}^p)$ and $\text{Herm}(\mathbb{C}^q)$ (from P_i), one gets $\dim X_{\lambda}^{p,q} = 2pq + p^2 + q^2 = (p+q)^2 = n^2$, which is as it should be.

Finally, why can one shift variables and do the Gaussian integral over Q (with translation-invariant DQ) by completing the square? This question is legitimate as the Schäfer–Wegner domain $X_{\lambda}^{p,q}$ lacks invariance under the required shift, which is $Q_i \mapsto Q_i - isz + \sum_j J_{ij} s M_j s$.

To complete the square in [12], introduce a parameter $t \in [0, 1]$ and consider the family of shifts

$$Q_i \mapsto Q_i + t(-isz + \sum_j J_{ij} s M_j s)$$

For fixed t , this shift takes $\mathcal{J} = (X_{\lambda}^{p,q})^{|\Lambda|}$ into another domain, $\mathcal{J}(t)$. Inspection shows that the function [12] still decreases rapidly (uniformly in the M_i) on $\mathcal{J}(t)$, as long as $t < 1$. Without changing the integral, one can add pieces to $\mathcal{J}(t)$ (for $t < 1$) at infinity to arrange for the chain $\mathcal{J} \leftarrow \mathcal{J}(t)$ to be a cycle. Because $\mathcal{J}(t)$ is homotopic to $\mathcal{J}(0) = \mathcal{J}$, this cycle is a boundary: there exists a manifold $\mathcal{V}(t)$ of dimension $\dim \mathcal{J} + 1$ such that $\partial \mathcal{V}(t) = \mathcal{J} - \mathcal{J}(t)$. Viewed as a holomorphic differential form of degree $(n^2|\Lambda|, 0)$ in the complex space $\text{End}(\mathbb{C}^n)^{|\Lambda|}$, the integrand $\omega := F_M(Q) DQ$ is closed (i.e., $d\omega = 0$). Therefore, by Stokes’ theorem,

$$\int_{\mathcal{J}} \omega - \int_{\mathcal{J}(t)} \omega = \int_{\partial \mathcal{V}(t)} \omega = \int_{\mathcal{V}(t)} d\omega = 0$$

which proves $\int_{\mathcal{J}(t)} F_M(Q) DQ = \int_{\mathcal{J}} F_M(Q) DQ$, independent of t . (This argument does not go through for the nonrigorous choice $sQ_i := T_i P_i T_i^{-1}$ usually made!)

In the limit $t \rightarrow 1$, one encounters the expression

$$\int_{\mathcal{J}(1)} F_M(Q) DQ = \int_{\mathcal{J}} d\nu_{n,J}(isQ) \times e^{-(1/2)\sum_{ij} J_{ij} \text{tr}(sM_i s M_j) + i\sum_k \text{tr}(sz M_k)}$$

with $d\nu_{n,J}$ as in [6]. The normalization integral over \mathcal{J} is defined by taking the Hermitian matrices P_i to be the inner variables of integration. The outer integrals over the T_i then demonstrably exist, and one can fix the (otherwise arbitrary) normalization of DQ by setting $\int_{\mathcal{J}} d\nu_{n,J}(isQ) = 1$. Making that choice, and comparing with [10], one has proved

$$\Omega_{n,N}^{\text{bos}} = \int_{\varphi, \bar{\varphi}} \left(\int_{\mathcal{J}} F_{(M_i)_{\alpha\beta} = (\bar{\varphi}_{\alpha}, \Pi_i \varphi_{\beta})}(Q) DQ \right)$$

The final step is to change the order of integration over the Q - and φ -variables, which is permitted since the Q -integral converges uniformly in φ . Doing the Gaussian φ -integral and shifting $Q_k \rightarrow Q_k - isz$, one arrives at the Schäfer–Wegner formula for $\Omega_{n,N}^{\text{bos}}$:

$$\Omega_{n,N}^{\text{bos}}(z, w^{-1}) = \int_{\mathcal{J}} e^{(1/2)\sum_{ij} w_{ij} \text{tr}(sQ_i s Q_j)} \times e^{-N \sum_k \text{tr} \ln(Q_k - isz)} DQ \quad [13]$$

which is a rigorous version of the naive formula [9]. Compared to Fyodorov’s formula, it has the disadvantage of not being manifestly invariant under global hyperbolic transformations $Q_i \mapsto T Q_i T^*$ (the integration domain \mathcal{J} is not invariant). Its best feature is that it does apply to the case of band random matrices with one orbital per site ($N = 1$).

Supersymmetric Variant

We are now in a position to tackle the problem of averaging ratios of determinants. For concreteness, we shall discuss the case where the number of determinants is two for both the numerator and the denominator, which is what is needed for the calculation of the function $G_{ij}^{(2)}(z_1, z_2)$ defined in eqn [2]. We will consider the case of relevance for the electrical conductivity: $z_1 = E + i\epsilon, z_2 = E - i\epsilon$, with $E \in \mathbb{R}$ and $\epsilon > 0$.

A Q -integral formula for $G_{ij}^{(2)}(z_1, z_2)$ can be derived by combining the fermionic method for

$$\left\langle \det(z_1 - H) \det(z_2 - H + t_2 E_{ij}^{ab}) \right\rangle$$

with the Schäfer–Wegner bosonic formalism for

$$\langle \det^{-1}(z_1 - H - t_1 E_{ji}^{ba}) \det^{-1}(z_2 - H) \rangle$$

and eventually differentiating with respect to t_1, t_2 at $t_1 = t_2 = 0$ and summing over a, b ; see the subsection “Green’s functions from determinants.” All steps are formally the same as before, but with traces and determinants replaced by their supersymmetric analogs. Having given a great many technical details in the last two sections, we now just present the final formula along with the necessary definitions and some indication of what are the new elements involved in the proof.

Let each of Q_{BB}, Q_{FF}, Q_{BF} , and Q_{FB} stand for a 2×2 matrix. If the first two matrices have commuting entries and the last two anticommuting ones, they combine to a 4×4 supermatrix:

$$Q = \begin{pmatrix} Q_{BB} & Q_{BF} \\ Q_{FB} & Q_{FF} \end{pmatrix}$$

Relevant operations on supermatrices are the supertrace,

$$\text{Str} Q = \text{tr} Q_{BB} - \text{tr} Q_{FF}$$

and the superdeterminant,

$$\text{Sdet} Q = \frac{\det(Q_{BB})}{\det(Q_{FF} - Q_{FB} Q_{BB}^{-1} Q_{BF})}$$

These are related by the identity $\text{Sdet} = \exp \circ \text{Str} \circ \ln$ whenever the superdeterminant exists and is nonzero.

In the process of applying the method described earlier, a supermatrix Q_i gets introduced at every site i of the lattice Λ . The domain of integration for each of the matrix blocks $(Q_i)_{BB} (i = 1, \dots, |\Lambda|)$ is taken to be the Schäfer–Wegner domain $X_{\lambda}^{1,1}$ (with some choice of $\lambda > 0$); the integration domain for each of the $(Q_i)_{FF}$ is the space of Hermitian 2×2 matrices, as before.

Let E_{BB}^{11} be the 4×4 (super)matrix with unit entry in the upper-left corner and zeros elsewhere; similarly, E_{FF}^{22} has unity in the lower-right corner and zeros elsewhere. Putting $s = \text{diag}(1, -1, 1, 1)$ and $z = \text{diag}(z_1, z_2, z_1, z_2)$, the supersymmetric Q -integral formula for the generating function of $G_{ij}^{(2)}$ – obtained by combining the Schäfer–Wegner bosonic method with the fermionic variant – is written as

$$\begin{aligned} & \left\langle \frac{\det(z_1 - H) \det(z_2 - H + t_2 E_{ij}^{ab})}{\det(z_1 - H - t_1 E_{ji}^{ba}) \det(z_2 - H)} \right\rangle \\ &= \int DQ e^{(1/2) \sum_{kl} w_{kl} \text{Str}(s Q_k s Q_l)} \\ & \times e^{-\text{Str} \ln (\sum_{r,c} (Q_r - isz) \otimes E_{rr}^{cc} + it_1 E_{BB}^{11} \otimes E_{ji}^{ba} - it_2 E_{FF}^{22} \otimes E_{ij}^{ab})} \end{aligned} \quad [14]$$

where the second supertrace includes a sum over sites and orbitals, and on setting $t_1 = t_2 = 0$ becomes

$$e^{-N \sum_r \text{Str} \ln (Q_r - isz)} = \prod_r \text{Sdet}^{-N} (Q_r - isz)$$

The superintegral “measure” $DQ = \prod_r DQ_r$ is the flat Berezin form, that is, the product of differentials for all the commuting matrix entries in $(Q_r)_{BB}$ and $(Q_r)_{FF}$, times the product of derivatives for all the anticommuting matrix entries in $(Q_r)_{BF}$ and $(Q_r)_{FB}$.

To prove the formula [14], two new tools are needed, a brief account of which is as follows.

Gaussian Superintegrals

There exists a supersymmetric generalization of the Gaussian integration formulas given in the subsection “Determinants as Gaussian integrals”: if $A, D(B, C)$ are linear operators or matrices with commuting (resp., anticommuting) entries, and $\Re A > 0$, one has

$$\text{Sdet}^{-1} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \int e^{-(\bar{\varphi}, A \varphi) - (\bar{\psi}, B \psi) - (\bar{\psi}, C \varphi) - (\bar{\varphi}, D \psi)}$$

Verification of this formula is straightforward. Using it, one writes the last factor in [14] as a Gaussian superintegral over four vectors: $\varphi_1, \varphi_2, \psi_1$, and ψ_2 . The integrand then becomes Gaussian in the matrices Q_r .

Shifting Variables

The next step in the proof is to do the “Gaussian” integral over the supermatrices Q_r . By definition, in a superintegral, one first carries out the Fermi integral, and afterwards the ordinary integrations. The Gaussian integral over the anticommuting parts $(Q_r)_{BF}$ and $(Q_r)_{FB}$ is readily done by completing the square and shifting variables using the fact that fermionic integration is differentiation:

$$\int d\xi f(\xi - \xi') = \frac{\partial}{\partial \xi} f(\xi - \xi') = \int d\xi f(\xi)$$

Similarly, the Gaussian integral over the Hermitian matrices $(Q_r)_{FF}$ is done by completing the square and shifting. The integral over $(Q_r)_{BB}$, however, is not Gaussian, as the domain is not \mathbb{R}^n but the Schäfer–Wegner domain. Here, more advanced calculus is required: these integrations are done by using a supersymmetric change-of-variables theorem due to Berezin to make the necessary shifts by nilpotents. (There is not enough space to describe this here, so please consult Berezin’s (1987) book.) Without difficulty, one finds the result to agree with the left-hand side of eqn [14], thereby establishing that formula.

Approximations

All manipulations so far have been exact and, in fact, rigorous (or can be made so with little extra effort). Now we turn to a sequence of approximations that have been used by physicists to develop a quantitative understanding of weakly disordered quantum dots, wires, films, etc. While physically satisfactory, not all of these approximations are under full mathematical control. We will briefly comment on their validity as we go along.

Saddle-Point Manifold

We continue to consider $G_{kl}^{(2)}(E + i\epsilon, E - i\epsilon)$ and focus on $E = 0$ (the center of the energy band) for simplicity. By varying the exponent on the right-hand side of [14] and setting the variation to zero one obtains, for $t_1 = t_2 = 0$,

$$\sum_i w_{ij} s Q_j s - N Q_i^{-1} = 0$$

which is called the saddle-point equation.

Let us now assume translational invariance, $w_{ij} = f(|i - j|)$. Then, if $\lambda = \sqrt{N/\Sigma_i w_{ij}}$, the saddle-point equation has i -independent solutions of the form

$$Q_i = \lambda \begin{pmatrix} q_{BB} & 0 \\ 0 & q_{FF} \end{pmatrix}$$

where for q_{FF} there are three possibilities: two isolated points $q_{FF} = \pm 1$ (unit matrix) coexist with a manifold

$$q_{FF} = \begin{pmatrix} \cos \theta_1 & \sin \theta_1 e^{i\phi_1} \\ \sin \theta_1 e^{-i\phi_1} & -\cos \theta_1 \end{pmatrix} \quad [15]$$

which is two-dimensional; whereas the solution space for q_{BB} consists of a single connected 2-manifold:

$$q_{BB} = \begin{pmatrix} \cosh \theta_0 & \sinh \theta_0 e^{i\phi_0} \\ \sinh \theta_0 e^{-i\phi_0} & \cosh \theta_0 \end{pmatrix} \quad [16]$$

The solutions $q_{FF} = \pm 1$ are usually discarded in the physics literature. (The argument is that they break supersymmetry and therefore get suppressed by fermionic zero modes. For the simpler case of the one-point function [1] and in three space dimensions, such suppression has recently been proved by Disertori, Pinson, and Spencer.) Other solutions for q_{BB} are ruled out by the requirement $\Re Q_i > 0$ for the Schäfer–Wegner domain.

The set of matrices [16] and [15] – the “saddle-point manifold” – is diffeomorphic to the product of a 2-hyperboloid H^2 with a 2-sphere S^2 . Moving

along that manifold $M := H^2 \times S^2$ leaves the Q -field integrand [14] unchanged (for $z_1 = z_2 = t_1 = t_2 = 0$).

One can actually anticipate the existence of such a manifold from the symmetries at hand. These are most transparent in the starting point of the formalism as given by the characteristic function $\langle e^{-iK_H} \rangle$ with

$$K_H = (\bar{\varphi}_1, H\varphi_1) - (\bar{\varphi}_2, H\varphi_2) + (\bar{\psi}_1, H\psi_1) + (\bar{\psi}_2, H\psi_2)$$

The signs of this quadratic expression are what is encoded in the signature matrix $s = \text{diag}(1, -1, 1, 1)$ (recall that the first two entries are forced by $\Im m z_1 > 0$ and $\Im m z_2 < 0$). The Hermitian form K_H is invariant under the product of two Lie groups: $U(1, 1)$ acting on the φ 's, and $U(2)$ acting on the ψ 's. This invariance gets transferred by the formalism to the Q -side; the saddle-point manifold M is in fact an “orbit” of the group action of $G := U(1, 1) \times U(2)$ on the Q -field. In the language of physics, the degrees of freedom of M correspond to the Goldstone bosons of a broken symmetry.

K_H also has some supersymmetries, mixing φ 's with ψ 's. At the infinitesimal level, these combine with the generators of G to give a Lie superalgebra of symmetries $\mathfrak{g} := \mathfrak{u}(1, 1|2)$. One therefore expects some kind of saddle-point supermanifold, say \mathcal{M} , on the Q -side.

\mathcal{M} can be constructed by extending the above solution $q_0 := \text{diag}(q_{BB}, q_{FF})$ of the dimensionless saddle-point equation $sqs = q^{-1}$ to the full 4×4 supermatrix space. Putting $q = q_0 + q_1$ with

$$q_1 = \begin{pmatrix} 0 & q_{BF} \\ q_{FB} & 0 \end{pmatrix}$$

and linearizing in q_1 , one gets

$$sq_1 s = -q_0^{-1} q_1 q_0^{-1} \quad [17]$$

The solution space of this linear equation for q_1 has dimension 4 for all $q_0 \in M$. Based on it, one expects four Goldstone fermions to emerge along with the four Goldstone bosons of M .

For the simple case under consideration, one can introduce local coordinates and push the analysis to nonlinear order, but things get quickly out of hand (when done in this way) for more challenging, higher-rank cases. Fortunately, there exists an alternative, coordinate-independent approach, as the mathematical object to be constructed is completely determined by symmetry!

Riemannian Symmetric Superspace

The linear equation [17] associates with every point $x \in M$ a four-dimensional vector space of solutions,

V_x . As the point x moves on M the vector spaces V_x turn and twist; thus, they form what is called a vector bundle V over M . (The bundle at hand turns out to be nontrivial, i.e., there exists no global choice of coordinates for it.)

A section of V is a smooth mapping $\nu: M \rightarrow V$ such that $\nu(x) \in V_x$ for all $x \in M$. The sections of V are to be multiplied in the exterior sense, as they represent anticommuting degrees of freedom; hence the proper object to consider is the exterior bundle, $\wedge V$.

It is a beautiful fact that there exists a unique action of the Lie superalgebra \mathfrak{g} on the sections of $\wedge V$ by first-order differential operators, or derivations for short. (Be advised however that this canonical \mathfrak{g} -action is not well known in physics or mathematics.)

The manifold M is a symmetric space, that is, a Riemannian manifold with G -invariant geometry. Its metric tensor, g , uniquely extends to a second-rank tensor field (still denoted by g) which maps pairs of derivations of $\wedge V$ to sections of $\wedge V$, and is invariant with respect to the \mathfrak{g} -action. This collection of objects – the symmetric space M , the exterior bundle $\wedge V$ over it, the action of the Lie superalgebra \mathfrak{g} on the sections of $\wedge V$, and the \mathfrak{g} -invariant second-rank tensor g – form what the author calls a “Riemannian symmetric superspace,” $\#$.

Nonlinear Sigma Model

According to the Landau–Ginzburg–Wilson (LGW) paradigm of the theory of phase transitions, the large-scale physics of a statistical mechanical system near criticality is expected to be controlled by an effective field theory for the long-wavelength excitations of the order parameter of the system.

Wegner is credited for the profound insight that the LGW paradigm applies to the random matrix situation at hand, with the role of the order parameter being taken by the matrix Q . He argued that transport observables (such as the electrical conductivity) are governed by slow spatial variations of the Q -field inside the saddle-point manifold. Efetov skilfully implemented this insight in a supersymmetric variant of Wegner’s method.

While the direct construction of the effective continuum field theory by gradient expansion of [14] is not an entirely easy task, the outcome of the calculation is predetermined by symmetry. On general grounds, the effective field theory has to be a nonlinear sigma model for the Goldstone bosons and fermions of $\#$: if $\{\phi^A\}$ are local coordinates for

the bundle V with metric $g_{AB}(\phi) = g(\partial/\partial\phi^A, \partial/\partial\phi^B)$, the action functional is

$$S = \sigma \int d^d x \partial_\mu \phi^A g_{AB}(\phi) \partial_\mu \phi^B$$

The coupling parameter σ has the physical meaning of bare (i.e., unrenormalized) conductivity. In the present model $\sigma = NW^2 a^{2-d}$, where W is essentially the width of the band random matrix in units of the lattice spacing a (the short-distance cutoff of the continuum field theory). S is the effective action in the limit $z_1 = z_2$. For a finite frequency $\omega = z_1 - z_2$, a symmetry-breaking term of the form $i\omega \nu \int d^d x f(\phi)$, where $\nu = N(\pi\lambda)^{-1} a^{-d}$ is the local density of states, has to be added to S .

By perturbative renormalization group analysis, that is, by integrating out the rapid field fluctuations, one finds for $d = 2$ that σ decreases on increasing the cutoff a . This property is referred to as “asymptotic freedom” in field theory. On its basis one expects exponentially decaying correlations, and hence localization of all states, in two dimensions. However, a mathematical proof of this conjecture is not currently available.

In three dimensions and for a sufficiently large bare conductivity, the renormalization flow goes toward the metallic fixed point ($\sigma \rightarrow \infty$), where G -symmetry is broken spontaneously. A rigorous proof of this important conjecture (existence of disordered metals in three space dimensions) is not available either.

Zero-Mode Approximation

For a system in a box of linear size L , the cost of exciting fluctuations in the sigma model field is estimated as the Thouless energy $E_{\text{Th}} = \sigma/\nu L^2$. In the limit of small frequency, $|\omega| \ll E_{\text{Th}}$, the physical behavior is dominated by the constant modes $\phi^A(x) = \phi^A$ (independent of x). By computing the integral over these modes, Efetov found the energy-level correlations in the small-frequency limit to be those of the GUE.

See also: Random Matrix Theory in Physics; Symmetry Classes in Random Matrix Theory.

Further Reading

- Berezin FA (1987) *Introduction to Superanalysis*. Dordrecht: Reidel.
- Disertori M, Pinson H, and Spencer T (2002) Density of states of random band matrices. *Communication in Mathematical Physics* 232: 83–124.
- Efetov KB (1997) *Supersymmetry in Disorder and Chaos*. Cambridge: Cambridge University Press.
- Fyodorov YV (2002) Negative moments of characteristic polynomials of random matrices: Ingham–Siegel integral as an

- alternative to Hubbard–Stratonovich transformation. *Nuclear Physics B* 621: 643–674.
- Mirlin AD (2000) Statistics of energy levels and eigen functions in disordered systems. *Physics Reports* 326: 260–382.
- Schäfer L and Wegner F (1980) Disordered system with n orbitals per site: Lagrange formulation, hyperbolic symmetry, and Goldstone modes. *Zeitschrift für Physik B* 38: 113–126.
- Wegner F (1979) The mobility edge problem: continuous symmetry and a conjecture. *Zeitschrift für Physik B* 35: 207–210.
- Zirnbauer MR (1996) Riemannian symmetric superspaces and their origin in random matrix theory. *Journal of Mathematical Physics* 37: 4986–5018.

Symmetric Hyperbolic Systems and Shock Waves

S Kichenassamy, Université de Reims
Champagne-Ardenne, Reims, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Many systems of partial differential equations arising in mathematical physics and differential geometry are quasilinear: the top-order derivatives enter only linearly. They may be cast in the form of first-order systems by introducing, if needed, derivatives of the unknowns as additional unknowns. For such systems, the theory of symmetric-hyperbolic (SH) systems provides a unified framework for proving the local existence of smooth solutions if the initial data are smooth. It is also convenient for constructing numerical schemes, and for studying shock waves. Despite what the name suggests, the impact of the theory of SH systems is not limited to hyperbolic problems, two examples being Tricomi's equation, and equations of Cauchy–Kowalewska type.

Application of the SH framework usually requires a preliminary reduction to SH form (“symmetrization”).

After comparing briefly the theory of SH systems with other functional-analytic approaches, we collect basic definitions and notation. We then present two general rules, for symmetrizing conservation laws and strictly hyperbolic equations, respectively. We next turn to special features possessed by linear SH systems, and give a general procedure to prove existence, which covers both linear and nonlinear systems. We then summarize those results on shock waves, and on blow-up singularities, which are related to SH structure. Examples and applications are collected in the last section.

The advantages of SH theory are: a standardized procedure for constructing solutions; the availability of standard numerical schemes; a natural way to prove that the speed of propagation of support is finite. On the other hand, the symmetrization process is sometimes *ad hoc*, and does not respect

the physical or geometric nature of the unknowns; to obviate this defect to some extent, we remark that symmetrizers may be viewed as introducing a new Riemannian metric on the space of unknowns. The search for a comprehensive criterion for identifying equations and boundary conditions compatible with SH structure is still the object of current research.

The most important fields of application of the theory today are general relativity and fluid dynamics, including magnetohydrodynamics.

Context of SH Theory in Modern Terms

The basic reason why the theory works may be summarized as follows for the modern reader; the history of the subject is, however, more involved.

Let H be a real Hilbert space. Consider a linear initial-value problem $du/dt + Au = 0$; $u(0) = u_0 \in H$, where A is unbounded, with domain $D(A)$. By Stone's theorem, one can solve it in a generalized sense, if the unbounded operator A satisfies $A + A^* = 0$. This condition contains two ingredients: a symmetry condition on A , and a maximality condition on $D(A)$, which incorporate boundary conditions (von Neumann, Friedrichs). Semigroup theory (Hille and Yosida, Phillips, and many others) handles more general operators A : it is possible to solve this problem in the form $u(t) = S(t)u_0$ for $t > 0$, where $\{S(t)\}_{t \geq 0}$ is a continuous contraction semigroup, if and only if $(Au, u) \geq 0$, and equation $x + Ax = y$ has a solution for every y in H (this is a maximality condition on $D(A)$). One then says that A is maximal monotone. For such operators, $A + A^* \geq 0$. SH systems are systems $Qu_t + Au = F$, satisfying two algebraic conditions ensuring formally that $A + A^*$ is bounded, and that Q is symmetric and positive definite. This algebraic structure enables one to solve the problem directly, without explicit reference to semigroup theory. Precise definitions are given next.

We assume throughout that all coefficients, nonlinearities, and data are smooth unless otherwise specified.

Definitions

Consider a quasilinear system

$$M_B^{A\alpha}(x, u) \partial_\alpha u^B = N^A(x, u) \quad [1]$$

where $u = (u^A)_{A=1, \dots, m}$, $x = (x^\alpha)_{\alpha=0, \dots, n}$, and $\partial_\alpha = \partial/\partial x^\alpha$. The components of u may be real or complex. We follow the summation convention on repeated indices in different positions; $x^0 = t$ may be thought of as the evolution variable; we write $x = (t, \mathbf{x})$, with $\mathbf{x} = (x^1, \dots, x^n)$. Indices A, B, \dots run from 1 to m , indices j, k, \dots from 1 to n , and Greek indices from 0 to n . The complex conjugate of u^A is written \bar{u}^A .

- Equation [1] is symmetrizable if there are functions $\sigma_{AB}(x, u)$ such that

$$M_{AB}^\alpha := \sigma_{AC} M_B^{C\alpha}$$

satisfies the condition $M_{AB}^\alpha = \bar{M}_{BA}^\alpha$ for every α .

- It is symmetric if it is symmetrizable with $\sigma_{AB} = \delta_{AB}$.
- It is symmetric-hyperbolic with respect to k_α if it is symmetric and if $k_\alpha M_{AB}^\alpha$ is positive definite: $k_\alpha M_{AB}^\alpha \xi^A \xi^B > 0$ for $\xi = (\xi^A) \neq 0$.

Thus, a symmetrizer (σ_{AB}) gives rise to a Riemannian metric $(k_\alpha \sigma_{AC} M_B^{C\alpha})$ on the space of unknowns, independent of any Riemannian structure on x -space. The system is SH with respect to x^0 if $k_\alpha = \delta_\alpha^0$.

The simplest class of SH systems is provided by real semilinear systems of the form

$$A^0(x) \partial_t u + A^i(x) \partial_i u = N(x, u) \quad [2]$$

where the A^α are real symmetric matrices, A^0 is symmetric and positive definite, and $k_\alpha = \delta_{\alpha 0}$. Writing $A^0 = P^2$, with P symmetric and positive definite, one finds that $v = Pu$ solves a SH system with $A^0 = I$ (identity matrix).

Conservation laws (with “reaction” or “source” term N^A) are usually defined as quasilinear systems of the form

$$\partial_\alpha f^{A\alpha}(x, u) = N^A(x, u) \quad [3]$$

They are common in fluid dynamics and combustion. They are limiting cases of nonlinear diffusion equations of the typical form

$$\partial_\alpha f^{A\alpha}(x, u) = N^A(x, u) + \varepsilon \partial_j (B_B^{Ajk} \partial_k u^B) \quad [4]$$

The determination of the form of the coefficients B_B^{Ajk} is a nontrivial modeling issue; they may reflect varied physical processes such as heat conduction, viscosity, or bulk viscosity. They may depend on x , u , and the derivatives of u . The simplest case is

$B_B^{Ajk} = D^{jk} \delta_B^A$ with (D^{jk}) diagonal. Some authors require the symmetry condition

$$\delta_{AC} B_B^{Cjk} = \delta_{BC} B_A^{Ckj} \quad [5]$$

Equations in which $f^{A\alpha} = u^A \delta_0^\alpha$ are called reaction–diffusion equations; they arise in physical and biological problems in which chemical reactions and diffusion phenomena are combined, and in population dynamics.

A conservation law is symmetric if and only if $\partial f^{A\alpha} / \partial u^B$ is symmetric in A and B , which means that there are, locally, functions $g^\alpha(x, u)$ such that $f^{A\alpha} = \partial g^\alpha / \partial u^A$.

A more fundamental derivation of conservation laws would take us beyond the scope of this survey.

Symmetrization

Two general procedures for symmetrization are available: one for conservation laws, the other for semilinear strictly hyperbolic problems.

Conservation Laws with a Convex Entropy

Consider, for simplicity, a conservation law of the form

$$\partial_t u^A + \partial_j f^{Aj}(u) = 0 \quad [6]$$

We, therefore, assume that the $f^{A\alpha} = f^{A\alpha}(u)$ and $f^{A0}(u) = u^A$. We show that the following three statements are equivalent locally: (1) there is a strictly convex function $U(u)$ such that $\sigma_{AB} = \partial^2 U / \partial u^A \partial u^B$ is a symmetrizer; (2) eqn [6] implies a scalar relation of the form $\partial_\alpha U^\alpha = 0$, with U^0 strictly convex; and (3) there is a change of unknowns $v_A = v_A(u)$ such that the system satisfied by $v = (v_A)$ is SH and $(\partial v_A / \partial u^B)$ is positive definite.

In fluid dynamics, U^0 may sometimes be related to specific entropy, and U^j to entropy flux. For this reason, if (2) holds, one says that U^0 is an entropy for eqn [6], and that (U^0, U^j) is an entropy pair. A system may have several entropies in this sense; this fact is sometimes useful in studying convergence properties of approximate solutions of eqn [6].

Let us now prove the equivalence of these properties.

Assume first (3): there are new unknowns $v_A = v_A(u)$ and functions $g^\alpha(v)$ such that $f^{A\alpha} = \partial g^\alpha / \partial v_A$. One finds that if eqn [6] holds,

$$\partial_\alpha U^\alpha = 0 \quad \text{where } U^\alpha = v_A \frac{\partial g^\alpha}{\partial v_A} - g^\alpha \quad [7]$$

Furthermore, we have $f^{A0} = u^A$; therefore, eqn [7] gives: $U^0 = v_A u^A - g^0$, so that U^0 is the Legendre

transform (familiar from mechanics) of g^0 . It follows that $v_A = \partial U^0 / \partial u^A$. Finally, $(\partial v_A / \partial u^B) = (\partial^2 U^0 / \partial u^A \partial u^B)$ is positive definite, and U^0 is strictly convex.

We have proved that (3) implies (2). Next, assume (2): the entropy equality $U_t + \partial_j U^j = 0$ holds identically – and not just for the solution at hand. Using [6], we find

$$\begin{aligned} 0 &= \frac{\partial U}{\partial u^A} \partial_t u^A + \frac{\partial U^j}{\partial u^B} \partial_j u^B \\ &= \left[-\frac{\partial U}{\partial u^A} \frac{\partial f^{Aj}}{\partial u^B} + \frac{\partial U^j}{\partial u^B} \right] \partial_j u^B \end{aligned}$$

Assumption (2), therefore, means that U is strictly convex and satisfies

$$\frac{\partial U}{\partial u^A} \frac{\partial f^{Aj}}{\partial u^B} = \frac{\partial U^j}{\partial u^B} \quad [8]$$

Now, letting $v_A = \partial U / \partial u^A$ and $g^j(v) = v_A f^{Aj} - U^j$, we find

$$\begin{aligned} \frac{\partial g^j}{\partial v_A} &= f^{Aj} + \left[\frac{\partial U}{\partial u^A} \frac{\partial f^{Aj}}{\partial u^C} - \frac{\partial U^j}{\partial u^C} \right] \frac{\partial u^C}{\partial v_A} \\ &= f^{Aj} \end{aligned} \quad [9]$$

Let $\sigma_{AB} = \partial^2 U / \partial u^A \partial u^B$. Since U is strictly convex, (σ_{AB}) is positive definite, and so is its inverse. We have now proved (3). Note that $u^A = \partial g^0 / \partial v_A$, where $g^0(v) = u^A v_A - U(u)$ is the Legendre transform of U .

Next, using eqn [9], and the relations $\sigma_{AB} = \partial v_A / \partial u^B = \partial v_B / \partial u^A$, we find

$$\begin{aligned} 0 &= \sigma_{AB} [\partial_t u^B + \partial_j f^{Bj}] \\ &= \sigma_{AB} \partial_t u^B + \frac{\partial v_B}{\partial u^A} \frac{\partial^2 g^j}{\partial v^B \partial u^C} \partial_j u^C \\ &= \sigma_{AB} \partial_t u^B + \frac{\partial^2 g^j}{\partial u^A \partial u^C} \partial_j u^C \end{aligned}$$

which is SH; therefore, σ_{AB} is a symmetrizer for eqn [6], and (1) is proved. Thus, (2) implies (1) and (3).

Finally, if (1) holds, $\sigma_{AC} \partial f^{Cj} / \partial u^B$ is symmetric in A and B . It follows that

$$\frac{\partial}{\partial u^C} \left[\frac{\partial U}{\partial u^A} \frac{\partial f^{Aj}}{\partial u^B} \right] = \sigma_{AC} \frac{\partial f^{Aj}}{\partial u^B} + \frac{\partial U}{\partial u^A} \frac{\partial^2 f^{Aj}}{\partial u^B \partial u^C}$$

is symmetric in B and C , so that there are, locally, functions U^j such that eqn [8] holds. Therefore, (U, U^j) is an entropy pair, and we see that (1) implies (2).

This completes the proof of the equivalence of (1), (2), and (3).

Strictly Hyperbolic Equations

Consider the scalar equation $Pf = g(t, x)$, where P is the linear operator

$$P = \partial_t^N - \sum_{j=0}^{N-1} p_{N-j}(t, x) \partial_t^j$$

of order N . Let $\Lambda = (1 - \Delta)^{1/2}$, where Δ is the Laplace operator on the space variables. Then $u = (u^A)$, where $u^A = \partial_t^{A-1} \Lambda^{N-A} f$ for $A = 1, \dots, N$, solves a first-order pseudodifferential system of the form

$$u_t - Lu = G$$

If P is strictly hyperbolic, the principal symbol $a_1(t, x, \xi)$ of L has a diagonal form with real eigenvalues $\lambda_j(t, x, \xi)$, and there are projectors $p_j(t, x, \xi) (p_j^2 = p_j)$ which commute with a_1 , such that $1 = \sum_j p_j$, and $a_1 = \sum_j \lambda_j p_j$. Let $r_0 = \sum_j p_j^* p_j$, and $r_0(D)$ the corresponding operator. Equation

$$r_0(D) \partial_t u - r_0(D) Lu = r_0(D) G$$

is formally SH in the following sense: r_0 is positive definite and $r_0 a_1$ is Hermitian.

Linear Problems

Consider a linear system

$$\begin{aligned} Lu &:= Q(t, x) \partial_t u + A^j(t, x) \partial_j u + B(t, x) u \\ &= f(t, x) \end{aligned} \quad [10]$$

We assume that Q and the A^j are real and symmetric, $Q \geq c$ with c positive, and all coefficients and their first-order derivatives are bounded.

Energy Identity

Multiplying the equation by u^T (transpose of u), one derives the “energy identity”

$$\partial_t (u^T Q u) + \partial_j (u^T A^j u) + u^T C u = 2u^T f(t, x) \quad [11]$$

where $C = 2B - \partial_t Q - \partial_j A^j$. C is not necessarily positive. However, $v := u \exp(-\lambda t)$ satisfies a linear SH system for which C is positive definite if λ is large enough.

Propagation of Support

A basic property of wave-like equations is finite speed of propagation of support: if the right-hand side vanishes, and if the solution at time 0 is localized in the ball of radius r , then the solution at time t is localized in the ball of radius $r + ct$ for a suitable constant c .

This property also holds for SH systems. To see this, let us consider the set where a solution u

vanishes: if the initial condition vanishes for $|\mathbf{x}| \leq R$, we claim that u at some later time vanishes for $|\mathbf{x}| \leq R - t/a$, for a large enough.

Indeed, let us integrate the energy identity on a truncated cone $\Gamma := \{|\mathbf{x}| \leq a(t_0 - t)/t_0; 0 \leq t \leq t_1\}$ with $t_1 < t_0$. The boundary of Γ consists of three parts: $\partial\Gamma = \Omega_0 \cup \Omega_1 \cup S$, where Ω_0 and Ω_1 represent the portions of the boundary on which $t=0$ and t_1 , respectively. The outer normal to S is proportional to $(a, t_0 \mathbf{x}^j/|\mathbf{x}|)$. Let $E(s)$ denote the integral of $u^T Q u$ on $\Gamma \cap \{t=s\}$. Integrating eqn [11] by parts, we obtain

$$\begin{aligned} E(t_1) - E(0) + \int_S u^T \Phi u \, ds \\ = \iint_{\Gamma} (2u^T f - u^T C u) \, dt \, d\mathbf{x} \end{aligned} \quad [12]$$

where Φ is proportional to $aQ + t_0 \sum_j x^j A^j/|\mathbf{x}|$. Take a so large that Φ is positive definite. The integral over S is then non-negative. If C is positive definite and $f \equiv 0$, so that $E(0)=0$, we find that $E(t_1) \leq 0$. Since Q is positive definite, this implies $u \equiv 0$ on Ω_1 , as claimed.

A Numerical Scheme

System $Lu=f$ may be discretized, for example, by the Lax–Friedrichs method: let h be the discretization step in space, and k the time step; write $\tau_j u(t, \mathbf{x}) = u(t, x^1, \dots, x^j + h, \dots, x^n)$ (translation in the j direction). One replaces $\partial_j u$ by the centered difference in the j direction: $(\tau_j u - \tau_j^{-1} u)/2h$; and the time derivative by

$$[u(t+k, \mathbf{x}) - \frac{1}{2n} \sum_j (\tau_j u(t, \mathbf{x}) + \tau_j^{-1} u(t, \mathbf{x}))]/k \quad [13]$$

For consistency of the scheme, we require $k/h = \lambda > 0$ to be fixed as k and h tend to zero; stability then holds if λ is small.

Nonlinear Problems and Singularities

We give a simple setup for proving the existence of smooth solutions to SH systems for small times. Such solutions may develop singularities. We limit ourselves to two types of singularities, on which SH structure provides some information: jump discontinuities and blow-up patterns. Caustic formation is not considered.

Construction of a Smooth Solution

Consider a real SH system (eqn [1]). Recall that a function of \mathbf{x} belongs to the Sobolev space H^s if its derivatives of order s or less are square-integrable.

One constructs a solution defined for t small, which is in H^s , $s > n/2 + 1$, as a function of \mathbf{x} , by the following procedure:

- (1) Replace spatial derivatives by regularized operators, which should be bounded in Sobolev spaces; the regularized equation is an ODE in H^s ; let u_ε be its solution.
- (2) Write the equation satisfied by derivatives of order s of u_ε , and apply the energy identity to it.
- (3) Find a positive T such that the solution is bounded in H^s for $|t| \leq T$, uniformly in ε ; this implies a C^1 bound.
- (4) Prove the convergence of the approximations in L^2 .
- (5) Prove the continuity in time of the H^s norm; conclude that the u_ε tend to a solution in $C(-T, T; H^s)$.

The result admits a local version, in which Sobolev spaces are replaced by Kato's "uniformly local" spaces. Uniqueness of the solution is proved along similar lines. We do not attempt to identify the infimum of the values of s for which the Cauchy problem is well-posed.

Jump Discontinuities: Shock Waves

A "shock wave" is a weak solution of a system of conservation laws admitting a jump discontinuity. By definition, weak solutions satisfy, for any smooth function $\phi_A(\mathbf{x})$ with compact support,

$$\iint \{f^{A\alpha} \partial_\alpha \phi_A + N^A \phi_A\} \, dt \, d\mathbf{x} = 0$$

The theory of shock waves is an attempt to understand solutions of conservation laws which are limits of solutions of diffusion equations; the hope is that the influence of second-derivative terms is appreciable only near shocks, and that, for given initial data, there is a unique weak solution of the conservation law which may be obtained as such a limit, if modeling has been done correctly. This problem may be difficult already for a single shock ("shock structure").

The theory of shock waves follows the one-dimensional theory closely. We therefore describe the main facts for a conservation law in one space dimension ($u = u(t, x)$):

$$\partial_t u + \partial_x f(u) = 0$$

If a shock travels at speed c , the weak formulation of the equations gives the Rankine–Hugoniot relation $c[u] = [f(u)]$, where square brackets denote jumps. There may be several weak solutions having the same initial condition. One restricts solutions by

making two further requirements: (1) the system admits an entropy pair (U, F) with a convex entropy and (2) to be admissible, weak solutions must be limits of “viscous approximations”

$$\partial_t u + \partial_x f(u) = \varepsilon \partial_x^2 u$$

as $\varepsilon \rightarrow 0$. One then finds easily that the entropy equality $(\partial_t U + \partial_x F = 0)$ must be replaced, for such weak solutions, by the entropy condition: $\partial_t U + \partial_x F \leq 0$ in the weak sense. This condition admits a concrete interpretation if the gradient of each characteristic speed is never orthogonal to the corresponding right eigenvector (“genuine nonlinearity”); in that case, characteristics must impinge on the shock (“shock inequalities”).

For the equations of gas dynamics with polytropic law $(pv^\gamma = \text{const.})$, there is a unique solution with initial condition $u = u_l$ for $x < 0, u = u_r$ for $x > 0$, where u_l and u_r are constant (“Riemann problem”) which satisfies the entropy condition, provided $|u_l - u_r|$ is small. More generally, if the equation of state $p = p(v, s) > 0$ satisfies $\partial p / \partial v < 0$ and $\partial^2 p / \partial v^2 > 0$, the shock inequalities are equivalent to the fact that the entropy increases after the passage of a shock with $|u_l - u_r|$ small.

On the numerical side, one should mention: (1) the widely used idea of upstream differencing; (2) the Lax–Wendroff scheme, the complete analysis of which requires tools from soliton theory; and (3) the availability of general results for dissipative schemes for SH systems.

Recent trends include: (1) admissibility conditions when genuine nonlinearity does not hold and (2) other approximations of shock wave problems, most notably kinetic formulations.

Some of the ideas of shock wave theory have been applied to Hamilton–Jacobi equations and to motion by mean curvature, with applications to front propagation problems and “computer vision.”

Stronger Singularities: Blow-Up Patterns

The amplitude of a solution may also grow without bound. Examples include optical pulse propagation in Kerr media and singularities in general relativity. The phenomenon is common when reaction terms are allowed. As we now explain, this phenomenon is reducible to SH theory in many cases of interest.

Blow-up singularities are usually not governed by the characteristic speeds defined by the principal part, because top-order derivatives are balanced by lower-order terms. In many applications, a systematic process (Fuchsian reduction) enables one to identify the correct model near blow-up; as a result,

one can write the solution as the sum of a singular part, known in closed form, and a regular part. If the singularity locus is represented by $t = 0$, the regular part solves a renormalized equation of the typical form

$$tMu + Au = t^\varepsilon N \quad [14]$$

where $Mu = 0$ is SH. Under natural conditions, for any initial condition u_0 such that $Au_0 = 0$, there is a unique solution of eqn [14] defined for small t .

The upshot is an asymptotic representation of solutions which renders the same services as an exact solution, and is valid precisely where numerical computation breaks down.

Fuchsian reduction enables one in particular to study (1) the blow-up time; (2) how the singularity locus varies when Cauchy data, prescribed in the smooth region, are varied; and (3) expressions which remain finite at blow-up. It is the only known general procedure for constructing analytically singular spacetimes involving arbitrary functions, rather than arbitrary parameters, and is therefore relevant to the search for alternatives to the big bang.

Examples and Applications

Wave Equation with Variable Coefficients

Consider the equation

$$\partial_{tt} u + 2a^j(x) \partial_{jt} u - a^{jk}(x) \partial_{jk} u = f(t, x, u, \nabla u)$$

with (a^{jk}) positive definite. Letting $v = (v_0, \dots, v_{n+1}) := (u, \partial_j u, \partial_t u)$, we find the system

$$\begin{aligned} \partial_t v_0 &= v_{n+1} \\ \partial_t v_k - \partial_k v_{n+1} &= 0 \\ \partial_t v_{n+1} + 2a^k \partial_k v_{n+1} - a^{jk} \partial_k v_j &= f \end{aligned}$$

It is symmetrizable, using the quadratic form $\sigma_{AB} v^A v^B = v_0^2 + a^{jk} v_j v_k + v_{n+1}^2$.

One proves directly that, if $v_j = \partial_j v_0$ for $t = 0$, this relation remains true for all t .

Maxwell’s Equations

Maxwell’s equations may be split into six evolution equations: $\partial_t E - \text{curl } B + j = 0$ and $\partial_t B + \text{curl } E = 0$, and two “constraints” $\text{div } E - \rho = 0, \text{div } B = 0$. The system of evolution equations is already in symmetric form; the quadratic form $\sigma_{AB} u^A u^B$ is here $|E|^2 + |B|^2$.

Compressible Fluids

Consider first the case of a polytropic gas:

$$\begin{aligned}\partial_t \mathbf{v} + (\mathbf{v} \cdot \nabla) \mathbf{v} + \rho^{-1} \nabla p &= 0 \\ \partial_t \rho + \operatorname{div}(\rho \mathbf{v}) &= 0\end{aligned}\quad [15]$$

with p proportional to ρ^γ . Taking (p, \mathbf{v}) as unknowns, one readily finds the SH system

$$\frac{1}{\gamma p} \partial_t p + \frac{1}{\gamma p} (\mathbf{v} \cdot \nabla) p + \operatorname{div} \mathbf{v} = 0 \quad [16]$$

$$\rho \partial_t \mathbf{v} + \nabla p + \rho (\mathbf{v} \cdot \nabla) \mathbf{v} = 0 \quad [17]$$

Symmetrization for more general compressible fluids with dissipation, including bulk viscosity, so as to satisfy the additional condition [5] may be achieved if we take as thermodynamic variables ρ and T , and assume pressure p and internal energy ε satisfy $\partial p / \partial \rho > 0$ and $\partial \varepsilon / \partial T > 0$, by taking as unknowns $(\rho, \rho \mathbf{v}, \rho(\varepsilon + |\mathbf{v}|^2/2))$. The specific entropy s satisfies $d\varepsilon = Tds - pd(1/\rho)$. If the viscosity and heat conduction coefficients are positive, one finds that $U = -\rho s$ is a convex entropy (in the sense of SH theory) on the set where $\rho > 0, T > 0$.

Einstein's Equations

The computation of solutions of Einstein's equations over long times, in particular in the study of coalescence of binary stars, has recently led to unexplained difficulties in the standard Arnowitt-Deser-Misner (ADM) formulation of the initial-value problem in general relativity. One way to tackle these difficulties is to rewrite the field equations in SH form; we focus on this particular aspect of recent research.

Recall the problem: find a four-dimensional metric g_{ab} with Lorentzian signature, such that $R_{ab} - \frac{1}{2} R g_{ab} = \chi T_{ab}$, with $\nabla^a T_{ab} = 0$, combined with an equation of state if necessary. R_{ab} is the Ricci tensor and $R = g^{ab} R_{ab}$ is the scalar curvature; they depend on derivatives of the metric up to order 2. In addition to the metric, T_{ab} involves physical quantities such as fluid 4-velocity or an electromagnetic field. The conservation laws of classical mathematical physics are all contained in the relation $\nabla^a T_{ab} = 0$.

Now, the field equations cannot be solved for $\partial_t^2 g_{ab}$, and, as a consequence, the Taylor series of g_{ab} with respect to time cannot be determined, even formally, from the values of g_{ab} and $\partial_t g_{ab}$ for $t=0$ (i.e., the Cauchy data). Furthermore, these data must satisfy four *constraint equations*. If the constraints are satisfied initially, they "propagate." But in numerical computation, these constraints are

never exactly satisfied, and the computed solution may deviate considerably from the exact solution. Also, numerical computations depend heavily on the way Einstein's equations are formulated.

The simplest way to derive a SH system is to replace R_{ab} by $R_{ab}^{(h)} = R_{ab} - \frac{1}{2} [g_{bc} \partial_a F^c + g_{ac} \partial_b F^c]$, where $F^c := g^{ab} \Gamma_{ab}^c$. It turns out that $R_{ab}^{(h)} = -\frac{1}{2} g^{cd} \partial_{cd} g_{ab} + H_{ab}(g, \partial g)$, where the expression of H_{ab} is immaterial. Applying to each component of the metric the treatment of the first example above (wave equation with variable coefficients), one easily derives an SH system of 50 equations for 50 unknowns: the ten independent components of the metric, and their 40 first-order derivatives. Now, if the Γ^c are initially zero (coordinates are "harmonic"), they remain so at later times.

Unfortunately, the harmonic coordinate condition does not seem to be stable in the large. More recent formulations start with one of the standard setups (ADM formalism, conformal equations, tetrad formalism, Newman-Penrose formalism) and proceed by adding combinations of the constraints to the equations, multiplied by parameters adjusted so as to ensure hyperbolicity or symmetric-hyperbolicity if needed. Another recent idea is to add a new unknown λ which monitors the failure of the constraint equations; one adds to the equations a new relation of the form $\partial_t \lambda = \alpha C - \beta \lambda$, where $C=0$ is equivalent to the constraints, and α and β are parameters. One then adds coupling terms to make the extended system SH. It is expected that the set of constraints acts as an attractor.

Reported computations indicate that these methods have resulted in an improvement of the time over which numerical computations are valid.

Tricomi's Equation

Let $\varphi(x, y)$ solve $(y \partial_x^2 - \partial_y^2) \varphi = 0$. Letting $u = e^{-\lambda x} (\partial_x \varphi, \partial_y \varphi)$, one finds a symmetric system $Lu = 0$, with

$$L = \begin{pmatrix} y & 0 \\ 0 & 1 \end{pmatrix} (\partial_x + \lambda) - \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \partial_y$$

If

$$Z = \begin{pmatrix} 1 & y \\ 1 & 1 \end{pmatrix}$$

we find that $K = ZL = A^1 \partial_x + A^2 \partial_y + B$, where

$$B - \frac{1}{2} (\partial_x A^1 + \partial_y A^2) = \begin{pmatrix} \frac{1}{2} + \lambda y & \lambda y \\ \lambda y & \lambda \end{pmatrix}$$

is positive definite if y is bounded, of arbitrary sign, and λ is small.

Cauchy–Kowalewska Systems

Consider a complex system

$$\partial_t u = A^i(z, t, u) \frac{\partial u}{\partial z^i} + B(z, t, u) \tag{18}$$

where $u = (u^A)$, $z = (z^1, \dots, z^n)$. The coefficients are analytic in their arguments when z and t are close to the origin and u is bounded by some constant K . The Cauchy–Kowalewska theorem ensures that, for any analytic initial condition near the origin, this system has a unique analytic solution near $z=0$, even without any symmetry assumption on the A^i . This result is a consequence of SH theory (Garabedian).

Indeed, write $z^j = x^j + iy^j$, $\partial_{z^j} = (1/2)(\partial_{x^j} - i\partial_{y^j})$, and $\partial_{\bar{z}^j} = (1/2)(\partial_{x^j} + i\partial_{y^j})$. Recall that analytic functions of z satisfy the Cauchy–Riemann equations $\partial_{\bar{z}^j} u = 0$.

Adding $(\bar{A}^i)^T \partial_{\bar{z}^i}$ to [18], and using the definition of ∂_{z^j} and $\partial_{\bar{z}^i}$, we find the symmetric system

$$u_t = \frac{1}{2} (A^i + (\bar{A}^i)^T) \partial_{x^i} u + \frac{1}{2i} (A^i - (\bar{A}^i)^T) \partial_{y^i} u + B \tag{19}$$

Solving this system, we find a candidate u for a solution of eqn [18]. To show that u is analytic if the data are, we solve a second SH system for $w = w^{(j)} := \partial_{\bar{z}^j} u$. If the data are analytic, w vanishes initially, and therefore remains zero for all t . Therefore, u is indeed analytic.

See also: Computational Methods in General Relativity: The Theory; Einstein Equations: Initial Value Formulation; Evolution Equations: Linear and Nonlinear; Magnetohydrodynamics; Partial Differential Equations: Some Examples; Semilinear Wave Equations; Shock

Wave Refinement of the Friedman–Robertson–Walker Metric.

Further Reading

Dautray R and Lions J (1988) *Mathematical Analysis and Numerical Methods for Science and Technology*, vols. 1–6. (transl. from French). New York: Springer.

Friedrichs KO (1986) In: Morawetz CS (ed.) *Selecta*, two volumes: collection of reprints with comments. Boston: Birkhäuser.

Godlewski E and Raviart P-A (1996) *Numerical Approximation of Hyperbolic Systems of Conservation Laws*. Berlin: Springer.

Gustafsson B, Kreiss H-O, and Olinger J (1996) *Time-Dependent Problems and Difference Methods*. New York: Wiley-Interscience.

John F (1982) *Partial Differential Equations*, 4th edn. Berlin: Springer.

Kato T (1975a) Quasilinear equations of evolution with applications to partial differential equations. *Lecture Notes in Mathematics* 448: 25–70.

Kato T (1975b) The Cauchy problem for quasilinear symmetric hyperbolic systems. *Archive for Rational Mechanics and Analysis* 58: 181–205.

Kawashima S and Shizuta Y (1988) On the normal form of the symmetric hyperbolic–parabolic systems associated with the conservation laws. *Tôhoku Mathematical Journal* 40: 449–464.

Kichenassamy S (1996) *Nonlinear Wave Equations*. New York: Dekker.

Lax PD (1973) *Hyperbolic Conservation Laws and the Mathematical Theory of Shock Waves. Regional Conference Series in Applied Mathematics*. Philadelphia: SIAM.

Lax PD and Phillips RS (1989) *Scattering Theory*, (revised edition). Boston: Academic Press.

Majda A (1984) *Compressible Fluid Flow and Systems of Conservation Laws*. Berlin: Springer.

Serre D (1999/2000) *Systems of Conservation Laws, I & II*. Cambridge: Cambridge University Press.

Smoller J (1983) *Shock Waves and Reaction-Diffusion Equations*. Berlin: Springer.

Taylor M (1991) *Pseudodifferential Operators and Nonlinear P.D.E., Progress in Mathematics*, vol. 100. Boston: Birkhäuser.

Symmetries and Conservation Laws

L H Ryder, University of Kent, Canterbury, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction: Spacetime Symmetries

Symmetries have played, and continue to play, an important role in fundamental physics, but the part they play is today seen as more complicated and many-sided than it was in the early days of particle physics, just after the Second World War. The area in which symmetries have had their most dramatic consequences is elementary particle physics, or

high-energy physics, and the majority of this article is concerned with this subject. The article concludes with some observations about symmetries and conservation laws in general relativity.

In the early days, considerations of symmetry were almost limited to Lorentz transformations: we begin by reviewing this crucially important topic. Invariance of the laws of nature under *translations* in space and time are actually necessary for the existence of science itself; if experiments did not yield the same results today and tomorrow, and in Paris and Moscow and on the Moon, then in effect there would be no laws of nature. Almost as strong

a statement could be made about invariance under *rotations*; if space were not isotropic, experimental results would depend on which direction the apparatus was aligned in, and again any laws would be extremely hard to find. Turning to the question of *motion*, Newton and Galileo realized that the laws of dynamics are the same in all inertial frames in relative motion. In the Newton–Galileo scheme, the rule for relating the space and time coordinates of two frames of reference is (for relative motion along the common x -axis)

$$x' = x - vt, \quad t' = t \quad [1]$$

This *principle of relativity* was reaffirmed by Einstein, but with the crucial modification that the rules for relating coordinates in two frames are given by Lorentz transformations, so that [1] is replaced by

$$x' = \gamma(x - vt), \quad t' = \gamma\left(t - \frac{vx}{c^2}\right) \quad [2]$$

Time is absolute in [1] but relative in [2]. Einstein was of course motivated by the fact that Maxwell's equations are covariant under Lorentz transformations, but not under Newton–Galileo ones.

The above considerations reveal that the laws of nature should be covariant under ten types of transformation: three translations in space, one in time, three parameters (angles) for rotations and three velocities. These transformations together form a group, the inhomogeneous Lorentz, or Poincaré group. It is a nonabelian group whose ten generators correspond to 4-momentum, angular momentum, and Lorentz boosts. The seminal work on the significance of this group in fundamental physics is that of Wigner in 1939. Assuming that the states of fundamental quantum systems (particles, atoms, molecules) form the basis states for *representations* of this group, these entities are described by two quantities, mass and spin. Spin, moreover, which was already familiar from earlier investigations in quantum physics, was described by the rotation group $SU(2)$, which is homomorphic to $SO(3)$ only for states with timelike momentum. For photons, for example, with null momentum, spin is described by the (noncompact) Euclidean group in the plane, with the consequence that there are only two polarization states for this massless particle.

Noether's theorem provides the crucial link between symmetries and *conservation laws*, via the principle of least action. Noether showed that the invariance of the action under a continuous symmetry operation implied the existence of a conserved quantity. The conserved quantities

corresponding to invariance under translation in space and time are momentum and energy; conservation of angular momentum follows from invariance under rotations and invariance under Lorentz transformations gives rise to conservation of motion of the center of mass.

Gauge Theories: Electromagnetism and Yang–Mills Theories

A quantity whose conservation has been well known for a long time is electric charge. The question may then be asked: invariance under what symmetry gives rise to conservation of electric charge? A classical complex field has the Lagrangian density

$$L = (\partial_\mu \phi)(\partial^\mu \phi^*) - m^2 \phi^* \phi \quad [3]$$

which is invariant under

$$\phi \rightarrow \exp(-iQ \Lambda) \phi \quad [4]$$

Λ being the parameter for the transformation. Noether's theorem then yields conservation of Q , interpreted as electric charge. With Λ a constant, as above, the Lagrangian possesses a “global” symmetry. This becomes a “local” symmetry when Λ becomes space and time dependent, $\Lambda(r, t)$ or $\Lambda(x^\mu)$. In that case, however, the Lagrangian [3] is no longer invariant under [4], because of the derivative terms. To preserve invariance an extra field A_μ must be introduced, so that [4] then becomes

$$\begin{aligned} \phi &\rightarrow \exp(-iQ \Lambda(x^\mu)) \phi \\ A_\mu &\rightarrow A_\mu + \frac{1}{Q} \partial_\mu \Lambda \end{aligned} \quad [5]$$

and the Lagrangian acquires extra terms, involving A_μ . The field A_μ is called a *gauge field* and is identified with the electromagnetic potential. The transformation [5] is called a *gauge transformation*, and since the phase factor $\exp(-iQ \Lambda)$ may be regarded as a unitary 1×1 matrix, we have here a theory with $U(1)$ gauge invariance, which describes electromagnetism and conservation of charge.

The notion of isospin had been introduced by Heisenberg in 1932. Isospin (then called isotopic spin) was a vector-like quantity conserved in strong (nuclear) interactions. Yang and Mills in 1954 made the pioneering suggestion that isospin conservation could also be recast as a gauge theory, by enlarging the $U(1)$ group of electromagnetism to $SU(2)$ (corresponding to rotations in “isospin space”), and at the same time treating the rotation angles as functions of spacetime. Then, eqn [4] will change: if

for example ψ is an isospinor field, then local isospin rotations are given by

$$\psi(x) \rightarrow \exp\left\{-i\frac{\tau}{2} \cdot \theta(x)\right\}\psi(x) = U(x)\psi(x) \quad [6]$$

where τ are the Pauli matrices: $\tau/2$ are the generators of SU(2). The gauge field then has three components $A_\mu^i (i=1,2,3)$ which may be written as a matrix

$$A_\mu = A_\mu^i \frac{\tau^i}{2}$$

transforming as

$$A_\mu \rightarrow A'_\mu = U(x)A_\mu U^{-1}(x) - \frac{i}{g}(\partial_\mu U(x))U^{-1}(x) \quad [7]$$

where g is the coupling constant, analogous to electric charge. The problem with this idea was that the isospin gauge field, analogous to the photon in electrodynamics, should, like the photon, be massless and have polarization states ± 1 (commonly, but inaccurately – see the work of Wigner (1939) – called spin 1); whereas the Yukawa particle, identified as the π meson, was massive and had spin 0, so could not act as the isospin gauge field.

The Yang–Mills idea really came into its own with the standard model (SM) of particle physics. This (gauge) model has an invariance group $SU(2) \otimes U(1) \otimes SU(3)$, the first two groups corresponding to electroweak interactions (a unification of weak interactions and electromagnetism) and the final SU(3) to quantum chromodynamics (QCD), the gauge theory describing quark interactions, which “glues” them together to make hadrons – protons, neutrons, pions, etc. This model is a dramatically successful one. The QCD sector of the theory requires essentially no further elaboration on the Yang–Mills idea than replacing the group SU(2) by SU(3). This is a straightforward matter of replacing the generators $\tau/2$ of SU(2) with the eight generators (3×3 matrices) of SU(3). $U(x)$ then also becomes a 3×3 matrix. The three degrees of freedom are the three quark “colors,” for which there is good experimental evidence, and the gluons, the quanta of the gauge fields, are indeed massless and have good experimental support. In the electroweak sector, however, the gauge fields, the W and Z bosons, were found with the predicted masses of 80.3 and 91.2 GeV respectively (the proton mass, for comparison, is 0.938 GeV). They are certainly not massless, as the straightforward Yang–Mills theory would require, and the explanation for this requires the introduction of the concept of spontaneous symmetry breaking.

Spontaneous Symmetry Breaking

The general idea of spontaneous symmetry breaking is that the *vacuum* – the state of lowest energy – is not invariant under the symmetry in question. A simple and common illustration is a pencil balanced vertically on its tip on a horizontal plane. The pencil is in unstable equilibrium but the system has a symmetry under rotations in the plane about the axis coincident with the pencil. Eventually, the pencil will fall into its lowest-energy state (vacuum), lying on the table in some direction – and the rotational symmetry is then lost. In fact, under rotations the actual lowest-energy (vacuum) state will be changed into another such state. There is a *degenerate vacuum*.

A similar scenario may be constructed in a complex scalar field theory. Consider such a theory with a Lagrangian given by

$$L = (\partial_\mu \phi)(\partial^\mu \phi^*) - m^2 \phi^* \phi - \lambda(\phi^* \phi)^2 \quad [8]$$

that is, with a potential energy function given by

$$V(\phi, \phi^*) = m^2 \phi^* \phi + \lambda(\phi^* \phi)^2 \quad [9]$$

where m is the mass of the field (quantum) and λ is the coupling of its self-interaction. The ground state is obtained by minimizing V , hence $\partial V / \partial \phi = 0$, giving (assuming that $m^2 > 0$) a minimum at $\phi = \phi^* = 0$. If, however, $m^2 < 0$, there is a local maximum at $\phi = 0$ and a minimum at $|\phi|^2 = -m^2 / 2\lambda > 0$. In quantum theory language, the vacuum expectation value $\langle 0 | \phi | 0 \rangle$ of the field is nonzero. Goldstone showed that this implied the presence of a massless scalar particle – a Goldstone boson. There was some interest in this result in particle physics, where the hypothesis of “partial conservation of the axial vector current” (PCAC) might result in a Goldstone boson that could be identified with the pion; although not massless, the pion is the lightest hadron, so “almost” massless.

Higgs analyzed what happens to the Goldstone model if electromagnetism is included. The Lagrangian [8] is invariant under the global transformation [4], but if this is made local, as in [5], a gauge field must be introduced and it is found that the massless Goldstone boson disappears and the massless gauge field (photon) becomes massive. Thus, spontaneous symmetry breaking of a gauge theory results in the appearance of a massive, rather than massless, gauge particle. (It is relevant to remark that a massless photon possesses two polarization states, but a massive one possesses three, so the number of spin-polarization states is preserved – the massless photon “eats” the Goldstone boson and becomes massive.) The Higgs model was generalized to the

case of a nonabelian symmetry group by Guralnik, Hagen, and Kibble and invoked by Weinberg in his 1971 model for the electroweak interaction in which the gauge quanta were massive.

Higgs' work was motivated by the theory of superconductivity, where the Meissner effect (expulsion of magnetic flux from a superconductor), when relativistic, implies that the effective mass of a photon in a superconductor is nonzero – this is, the “reason” that the flux does not penetrate. In the theory of Bardeen, Cooper, and Schrieffer (BCS), a superconductor is described by an effective scalar field, a composite of electron pairs (though paired in momentum space rather than coordinate space), and this provides a physical analogy with the model above. The SM of particle physics postulates a Higgs scalar field analogous to the BCS composite scalar field. If this field exists, Higgs particles should also exist, but they have not yet been found. This is an outstanding problem for the SM.

Baryon and Lepton Numbers

The fact that the proton p does not decay into positron plus photon, $e^+ + \gamma$, or muon plus photon, $\mu^+ + \gamma$, implies a conservation law of baryon number B (the proton possessing $B=1$ and the others $B=0$). Furthermore, the stability of μ^- and τ^- against decay into $e^- + \gamma$ implies conservation of lepton numbers L_e , L_μ , and L_τ . These are regarded as global, not local, symmetries, so there are no associated gauge fields or interactions. Interestingly, however, these symmetries are not built into the SM, so are not guaranteed by it. More interestingly, these symmetries are actually destroyed in one attempt to go beyond the SM. This is the hypothesis that QCD may be unified with electroweak interactions to produce a “grand unified” theory (GUT). The simplest GUT is the one in which the $SU(2) \otimes U(1) \otimes SU(3)$ symmetry is assumed to be a subgroup of the much tighter symmetry $SU(5)$, and in that theory the proton is unstable:

$$p \rightarrow e^+ + \pi^0 \quad [10]$$

The predicted lifetime is $10^{30 \pm 1}$ years, while a recent estimate of the lifetime for this decay mode is $> 5 \times 10^{32}$ years. It may be that GUTs do not exist in nature, but since the decay [10] violates conservation of the quantities B and L_e , even entertaining the idea that the decay might take place begs the question, “are these conservation laws sacrosanct?”

Another recent development which leads to the same question is the subject of neutrino oscillations. A strong motivation for this is the solar neutrino

problem; this is the problem that the number of electron neutrinos detected on Earth, originating in the Sun, is less than the number predicted, by a factor close to 3. The mismatch could be at least partly, and perhaps completely, explained if electron neutrinos “oscillated” into muon and/or tau neutrinos on their passage from the Sun to the Earth, since the reaction which detects the neutrinos on Earth is sensitive only to electron neutrinos, and not to the other species. But oscillation is only permitted if L_e , L_μ , and L_τ are not separately conserved quantities. Oscillation can also only take place if the masses of the different neutrinos are different – the oscillation rate depends on Δm^2 – hence not all the neutrinos may be massless.

Discrete Symmetries

Ever since parity violation was discovered in weak interactions (nuclear beta decay) by Wu in 1957, the whole subject of discrete symmetries has presented problems which are still not resolved. The symmetries in question are

P (space inversion): $(x, y, z) \rightarrow (-x, -y, -z)$

T (time reversal): $t \rightarrow -t$

C (particle–antiparticle conjugation): particle \leftrightarrow antiparticle

Are the laws of physics invariant under these operations? The Wu experiment revealed that weak interactions are not invariant under P , but what about other interactions and other operations? In this context, the CPT theorem is highly important. According to this theorem (based on very general assumptions), all laws of nature must be invariant under the combined operation CPT , so that, for example, the fact that weak interactions are not invariant under P means that they are not invariant under the product CT either.

The violation of P invariance in beta decay was soon related to the fact that the neutrino involved (the electron neutrino – or, to be precise, antineutrino) was massless. Spin-1/2 particles like the electron and neutrino obey the Dirac equation, which may be written out as a pair of coupled equations for left- and right-handed states. In the case $m=0$, however, these equations decouple so it is possible to have a massless spin-1/2 particle which is either left-handed or right-handed. Any interaction involving this particle would automatically violate parity (which turns a left-handed state into a right-handed one). Experiments have verified that the neutrino is indeed left-handed. The SM incorporates this in the sense that the left-handed electron e^-_L and the electron neutrino ν_e are assigned to a

weak isospin $SU(2)$ doublet, while the right-handed electron e^-_R transforms as a singlet. A similar pattern is repeated for the μ and τ particles and their neutrinos. The phenomenon of neutrino oscillations, on the other hand, does not allow all the neutrino states also to be purely left-handed (since they cannot be massless). This poses a potential problem for the SM.

For a few years after 1957 it was believed that beta decay violated C as well as P , but conserved the product CP ; and indeed that all weak interactions were CP invariant. In 1964, however, it was found that there is a small element of CP violation in K^0 decay. CP -violating effects are also expected in B^0 decays. The physical origin of CP violation is still not understood, but its importance is that it implies T violation, so that in (at least some) weak interactions, there is an “arrow of time” on the subnuclear scale. (Such an arrow of time is, of course, familiar in thermodynamics.) This is used in a cosmological context to explain baryon–antibaryon asymmetry in the Universe.

Baryon–Antibaryon Asymmetry

In the standard model of cosmology it is shown that applying the known laws of physics to the early Universe (the first few minutes) leads to the conclusion that at an age of 226 s nuclear fusion reactions took place resulting in a mixture of 74% protons and 26% α particles, so that, hundreds of thousands of years later, when galactic condensation took place, it would involve precisely this admixture of hydrogen and helium gases. Just this amount of helium has been found in the Sun, giving great confidence to the “big bang” model. Assuming that at extremely small times the baryon number of the Universe was zero, $B=0$, and assuming also (a big assumption, but one nevertheless made by cosmologists) that the Universe is made of matter and not antimatter, we may then ask, why is this – where has the antimatter gone?

Surprisingly, this question was addressed as early as 1966 by Sakharov, who showed that, starting with an initial state with $B=0$, it would be possible to reach a state with $B \neq 0$ as long as three conditions obtained: B violating interactions, CP and C violating interactions, and lack of thermal equilibrium. GUTs and ordinary weak interactions already provide possibilities for the first two of these conditions. Breakdown of thermal equilibrium will be expected to occur as the Universe expands. When the particle density is high, reactions such as $p + \bar{p} \rightarrow \gamma + \gamma$ will ensure an equal population of baryons and antibaryons, even in the presence of B

violating interactions, but as the density increases and this reaction rate becomes less than the expansion rate, thermal equilibrium can no longer be maintained. Thus, GUTs offer an explanation of why there is no antimatter in the Universe. It might be thought that this sort of explanation is implausible, since the B -violating and CP -violating forces are so weak, but actually this is not a problem, since the ratio of baryon number to photon number in the Universe is of the order $N_B/N_\gamma \approx 10^{-9}$; so we may conjure up a scenario in which the B and CP violating forces give rise to a volume of space in which there are, say, 10^9 antibaryons, $10^9 + 1$ baryons and approximately the same number of photons. Then, all the antibaryons become annihilated leaving one baryon and 10^9 photons – as observed.

A recent development in the area of discrete symmetries has been the suggestion by Kostelecky and coworkers that there might exist spontaneous violation of CPT and Lorentz symmetry.

Topological Charges

Conserved quantities of a quite different type have received a lot of attention in recent decades. Their conservation is a consequence of nontrivial boundary conditions for the fields. A famous example is the sine-Gordon “kink.” The sine-Gordon equation

$$\frac{\partial^2 \phi}{\partial t^2} - \frac{\partial^2 \phi}{\partial x^2} + \frac{1}{b^2} \sin(b\phi) = 0 \quad [11]$$

describes a scalar field in one space and one time dimension. It is a nonlinear equation which possesses, among others, the interesting solution

$$f(\xi) = \frac{4}{b} \arctan \exp[\pm(\gamma/\sqrt{b})\xi]$$

where $\xi = x - vt$ and $\gamma = (1 - v^2)^{-1/2}$. This corresponds to a solitary wave which moves, preserving its shape and size – in distinction to usual waves, which spread out and dissipate. Waves of this type are called solitons, and solitons have in fact been observed moving along canals. In this case, they are solutions to the Korteweg de Vries equation. Equation [11] clearly possesses the constant solutions

$$\phi = \frac{2\pi n}{b}, \quad n = 0, \pm 1, \pm 2, \dots$$

which, it may be shown, all have zero energy. We may then construct a solution of the above type, but with $n=0$ as $x \rightarrow -\infty$ and $n=N$ as $x \rightarrow +\infty$. This so-called “kink” solution has finite energy and is not continuously deformable into a solution with $n=0$ everywhere, since this would involve overcoming an

infinite energy barrier. The “kink number” may be characterized as a charge: defining the current

$$J^\mu = \frac{b}{2\pi} \varepsilon^{\mu\nu} \partial_\nu \phi$$

with $\varepsilon^{\mu\nu}$ the totally antisymmetric symbol, it is clear that this is identically conserved, $\partial_\mu J^\mu = 0$. This is a consequence of the definition of $\varepsilon^{\mu\nu}$; it is not a consequence of invariance of the sine-Gordon Lagrangian under a symmetry operation, so the current J^μ is not a Noether current. The associated conserved charge is

$$\begin{aligned} Q &= \int J^0 dx = \frac{b}{2\pi} \int \frac{\partial \phi}{\partial x} dx \\ &= \frac{b}{2\pi} [\phi(\infty) - \phi(-\infty)] = N \end{aligned}$$

Models of the above type may be written down in a spacetime with more than two dimensions. In that case the above solution depends only on one coordinate, so represents an infinite planar “domain wall,” on the two sides of which the field assumes different values. Such domain walls, as well as “cosmic strings,” are considered as serious possibilities in cosmology.

Nonabelian gauge theories and the sigma model also provide a fertile ground for topological excitations – field configurations which for topological reasons do not decay. Gauge theories with spontaneous symmetry breaking have two-dimensional solutions corresponding to vortex lines and three-dimensional solutions corresponding to magnetic monopoles. In spacetime (3 + 1 dimensions), there is a solution to the gauge field equations, with no spontaneous symmetry breaking, corresponding to an “instanton,” a finite-energy field configuration, localized in time as well as in space (hence the name). The gauge group here is SU(2), whose group space is S^3 . Spacetime is “Euclideanized” into R^4 , whose boundary is then S^3 . Asymptotic field configurations may then be characterized by mappings of S^3 in field space into S^3 in parameter space, and since the third homotopy group of S^3 is nontrivial, $\pi_3(S^3) = \mathbb{Z}$, these field configurations belong to different classes and are not deformable into each other. These define “degenerate vacua” of the gauge field equations. In quantum theory, tunneling between these vacua is allowed and ’t Hooft has shown how this may give rise to deuteron decay $d \rightarrow e^+ + \bar{\nu}_\mu$. Other examples of topologically nontrivial configurations are so-called sphalerons, which may also contribute to baryon number violation in the early Universe, and skyrmions, constructs in the nonlinear sigma model which serve as a model for baryon number.

Supersymmetry

Supersymmetry is a fermion–boson symmetry, postulating that multiplets of fundamental particles contain both fermions and bosons. Thus, for example, since electrons exist there should also be “selectrons” – “scalar” electrons, with spin 0. There should also be photinos, with spin 1/2, to take their place alongside photons, and so on. If supersymmetry were exact, these particles would have the same mass as their partners and would have all been found, but in fact none have yet been discovered, so presumably supersymmetry is a broken symmetry. The feature that makes supersymmetry attractive is that it holds some promise for solving divergence problems in quantum field theory, since the radiative corrections from fermion and boson loops are opposite in sign and may exactly cancel. Supersymmetric models can also help to solve the so-called hierarchy problem in quantum field theory.

If supersymmetry is made into a local symmetry, rather than simply a global one, extra fields must be introduced (as the photon field was introduced above), and it turns out that one of these is a spin-2 field, which may be identified with the *graviton*. Local supersymmetry thus becomes supergravity.

General Relativity

Symmetries and conservation laws take on new aspects when general relativity is considered. Einstein’s field equations relate the energy–momentum tensor of matter (and radiation) to the Ricci tensor of spacetime. The Ricci tensor has vanishing *covariant* divergence, which means that the energy–momentum tensor possesses the same property, but conservation of energy and momentum requires that it is the ordinary derivative, not the covariant one, of this tensor that should vanish. It might be expected that this problem could be alleviated by including the contribution of the gravitational field itself in energy–momentum tensor. This is quite reasonable, but then problems of interpretation arise, since at any one point in a general spacetime, a coordinate system might be found which is inertial (this is the force of the equivalence principle), corresponding to no gravitational field, and therefore no energy. The usual procedure is to introduce an energy–momentum “pseudotensor,” and to conclude that energy in a gravitational field is not localizable.

The role of symmetries in general relativity is rather different from its role in particle physics, which is set in Minkowski spacetime. In a general spacetime there are no symmetries, but many examples of particular spacetimes with their own symmetries are now known. The symmetry operations involved are

isometries, with corresponding groups of motion (so that the isometry group of Minkowski space is the Poincaré group). These groups are an important subject of study in cosmology; for example, there is a classification of homogeneous cosmological models, labeled according to the Bianchi classification.

See also: Cotangent Bundle Reduction; Effective Field Theories; Electroweak Theory; General Relativity: Overview; Infinite-Dimensional Hamiltonian Systems; Noncommutative Geometry and the Standard Model; Quantum Field Theory: A Brief Introduction; Quasiperiodic Systems; Sine-Gordon Equation; Supergravity; Symmetries in Quantum Field Theory of Lower Spacetime dimensions; Symmetry and Symplectic Reduction; Symmetry Classes in Random Matrix Theory; Topological Defects and Their Homotopy Classification.

Further Reading

- Aitchison IJ and Hey AJ (1981) *Gauge Theories in Particle Physics*. Bristol: Adam Hilger.
 Cheng T-P and Li L-F (1984) *Gauge Theory of Elementary Particle Physics*. Oxford: Clarendon Press.

- Eguchi T, Gilkey PB, and Hanson AJ (1980) Gravitation, gauge theories and differential geometry. *Physics Reports* 66: 213.
 Huang K (1998) *Quantum Field Theory: From Operators to Path Integrals*. New York: Wiley.
 Kosteletzky VA (ed.) (2004) *CPT and Lorentz Symmetry: Proceedings of the Third Meeting*. Singapore: World Scientific.
 Landau LD and Lifshitz EM (1971) *The Classical Theory of Fields*. Oxford: Pergamon Press.
 Manton N and Sutcliffe P (2004) *Topological Solitons*. Cambridge: Cambridge University Press.
 Perkins DH (2000) *Introduction to High Energy Physics*, 4th edn. Cambridge: Cambridge University Press.
 Review of Particle Properties (2002), *Physical Review D* 66: 01002, July 2002, Part 1.
 Rubakov V (2002) *Classical Theory of Gauge Fields*. Princeton: Princeton University Press.
 Ryder LH (1996) *Quantum Field Theory*, 2nd edn. Cambridge: Cambridge University Press.
 Stephani H (2004) *Relativity: An Introduction to Special and General Relativity*, 3rd edn. Cambridge: Cambridge University Press.
 Weinberg S (1983) *The First Three Minutes*. London: Fontana.
 Wess J and Bagger (1983) *Supersymmetry and Supergravity*. Princeton: Princeton University Press.
 Wigner E (1939) On unitary representations of the inhomogeneous Lorentz group. *Annals of Mathematics* 40: 149.

Symmetries in Quantum Field Theory of Lower Spacetime Dimensions

J Mund, Universidade de São Paulo, São Paulo, Brazil
K-H Rehren, Universität Göttingen, Göttingen, Germany

© 2006 Elsevier Ltd. All rights reserved.

Symmetries in Quantum Field Theory

Symmetries have proved to be one of the most powerful concepts in quantum theory, and in quantum field theory in particular. From the beginnings of quantum mechanics, it is well known that the presence of a symmetry allows one to predict relations between different measurements, to classify spectra (energy or other), and to understand the Pauli exclusion principle, to name only a few applications. Much more remarkably, in modern relativistic quantum field theory, designed to describe the interactions of elementary particles, fundamental interactions have been found to be induced by the principle of local gauge invariance.

One distinguishes spacetime symmetries (Poincaré or conformal transformations), which change the position and orientation of the system in space and time, and internal symmetries, which preserve the localization, acting on certain internal degrees of

freedom. The Coleman–Mandula (1967) theorem states that internal and spacetime symmetries cannot be mixed, in the sense that the generators of internal symmetries must be Lorentz scalars, hence the total group of symmetries factorizes into a direct product. Supersymmetries are an exception of this theorem because their generators do not form a Lie algebra, and they were in fact designed to circumvent the Coleman–Mandula theorem.

It is well known that the structure of symmetries of quantum systems in low-dimensional spacetime differs significantly from that in four-dimensional spacetime. (“Low” means in our context two or three, depending on the type of charge localization, c.f. below.) To name some examples:

- Two-dimensional quantum systems may have much higher symmetries than four-dimensional ones:
 - In two dimensions, there exist massive integrable models with infinitely many conservation laws and factorizable scattering matrices (*see* Integrability and Quantum Field Theory). These models exhibit solitonic superselection sectors, c.f. below.
 - The conformal group of two-dimensional spacetime is infinite dimensional, allowing for

the exact computation of correlation functions by the help of Ward identities (Belavin, Polyakov, and Zamolodchikov 1984). Only the finite-dimensional Möbius group, however, is also a symmetry of the vacuum state. Möbius covariance implies that the theory contains two subtheories of chiral fields defined on the light rays $t - x = \text{constant}$, resp. $t + x = \text{constant}$, and that these can be extended to fields defined on a circle, by adding a “point at infinity” to the light ray (Lüscher and Mack 1976). One arrives thus at one-dimensional chiral quantum field theories on a circle, which will play an important role in the discussion below.

- Continuous symmetries cannot be spontaneously broken in two dimensions. The latter is true not only for relativistic quantum field theory (Coleman 1973), but also in quantum statistical mechanics (Mermin and Wagner 1966) where it is responsible for the absence of ferromagnetism (see Symmetry Breaking in Field Theory). Spontaneous symmetry breakdown requires long-range order which is overcome by thermal fluctuations down to zero temperature, because these diverge logarithmically (in the thermodynamical limit) in two dimensions. This theorem thus illustrates how the spacetime dimension-dependent size of phase space has an effect on internal symmetries of quantum systems. A detailed mathematical analysis of the balance between phase space (thermal fluctuations) and long-range order (symmetry breakdown) has been given in a recent discussion of the Goldstone theorem (Buchholz, Doplicher, Longo and Roberts 1992).
- The Coleman–Mandula theorem, excluding a mixing between internal and spacetime symmetries (see above), is valid only in higher dimensions.

In more recent times, it has become apparent that low-dimensional quantum systems do not only admit more symmetries, but they may exhibit internal symmetries of an entirely new type, not describable by groups of transformations. In this article, we shall focus on the various ways in which the new symmetries can arise, and how they can be understood. In order to properly appreciate these issues, let us first recall some basic symmetry concepts in the conventional case.

In the traditional setting, symmetries arise in the form of groups of transformations of the quantum system which leave observable quantities (e.g., vacuum expectation values and correlation

functions) invariant. The symmetries form a group of $*$ -automorphisms of the algebra of fields:

$$\begin{aligned}\alpha_g(\phi_1\phi_2) &= \alpha_g(\phi_1)\alpha_g(\phi_2) \\ (\alpha_g(\phi))^* &= \alpha_g(\phi^*) \\ \alpha_{g_1}\alpha_{g_2} &= \alpha_{g_1g_2}\end{aligned}\quad [1]$$

(typically given by linear transformations of field multiplets). In the strongest case, the automorphisms are implemented by unitary operators on the state space

$$U(g)\phi U(g)^* = \alpha_g(\phi) \quad [2]$$

The implementers form a representation of the group of automorphisms,

$$U(g_1)U(g_2) = U(g_1g_2) \quad [3]$$

and there is an invariant vector state (a ground state, or the vacuum state in relativistic quantum field theory),

$$U(g)\Omega = \Omega \quad [4]$$

However, depending on the dynamics of the quantum system, these relations cannot always be fully realized. One therefore considers several weaker or more general notions of symmetries relevant in four dimensions:

- *Spontaneously broken symmetries.* The transformations are given as automorphisms of an algebra, but which are not unitarily implemented in a given irreducible representation of the algebra. Invariant pure states do not exist.
- *Projective representations.* The symmetries are unitarily implemented, but the implementers fail to satisfy the group law [3]. They give rise to ray (projective) representations or representations of a covering group. In particular, an invariant state vector as in [4] cannot exist in an irreducible representation.
- *Infinitesimal symmetries.* Lie algebras of infinitesimal transformations, given as derivations of an algebra, which cannot be integrated to finite transformations. Derivations may or may not be implemented in a given representation of the algebra by commutators with self-adjoint generators.
- *Supersymmetry.* The infinitesimal transformations form a graded Lie algebra.
- *Local gauge symmetries* form an infinite-dimensional group which are, however, not realized as automorphisms of the quantum algebra. Quantization of classical gauge interactions usually proceeds by breaking the gauge invariance in some way and restoring it at a later stage.

The Connection between Symmetry and Superselection Sectors

It is often convenient to describe a model in terms of localized fields which do not represent an observable (in the sense of quantum mechanics that an operator corresponds to some measurement prescription). For example, Fermi fields which violate the principle of causality because they anticommute with each other at spacelike distance rather than commute are not observables. Only fields which are quadratic in the Fermi fields (densities of charge, current, energy) are observables. This means that an internal symmetry is used in order to distinguish the observables as those operators which are invariant under the symmetry: in the example, the symmetry transformation multiplies each Fermi field by -1 (by the spin-statistics theorem, this transformation coincides with the univalence of the Lorentz group). We characterize this situation by writing

$$A(O) = F(O)^G \quad [5]$$

where $A(O)$ and $F(O)$ stand for the algebras of observables and fields localized in some spacetime region O , respectively, G is the internal symmetry group acting by automorphisms on each $F(O)$ without affecting the localization, and $F(O)^G \equiv \{a \in F(O), \alpha_g(a) = a \text{ for all } g \in G\}$ denotes the subalgebra of invariants. The internal symmetry group G which distinguishes the observables according to [5] is usually called the “(global) gauge group.”

If the gauge symmetry G is unbroken in the vacuum state, then there is a well-known connection between symmetry and superselection rules (see Symmetries and Conservation Laws): namely, the observables act reducibly on the vacuum Hilbert space representation of F because they commute with the unitary operators which implement the symmetry (or with their infinitesimal generators, usually called charges). As a consequence, the validity of the superposition principle is restricted because two eigenstates of different eigenvalues of the charges cannot exhibit interference. In other words, they belong to different superselection sectors. Wick, Wightman, and Wigner (1952) were the first to point out this relation. We therefore call this scenario the “WWW scenario” for brevity.

In the WWW scenario, the decomposition of the Hilbert space is determined by the central decomposition of the internal symmetry group (the eigenvalues of the Casimir operators). In this way, the superselection sectors are in one-to-one correspondence with the irreducible representations of the internal symmetry group.

Superselection sectors of two-dimensional models do not follow this scheme expected by the WWW scenario (see below). This was most strikingly demonstrated through the classification of the unitary highest-weight representations of the Virasoro algebra (Friedan, Qiu, and Shenker) which is nothing other than the classification of the superselection sectors of the observable algebra generated by the chiral stress-energy tensor, and through the determination of their fusion rules by Belavin, Polyakov, and Zamolodchikov (1984).

In two dimensions, one is therefore lacking a compelling *a priori* ansatz, like the WWW scenario, for describing the system in terms of auxiliary nonobservable charged fields. At this point, one may argue that from an operational point of view, a quantum field theory, and in particular its symmetries, should be understood entirely in terms of its observables. (This viewpoint is emphasized in the algebraic approach to QFT, see Algebraic Approach to Quantum Field Theory.) We shall therefore now ask the opposite question: suppose we are given an algebra A of local observables (without knowledge of a field algebra and its gauge group). We define the superselection sectors intrinsically as (the unitary equivalence classes of) the positive-energy representations of A . Then the question is: do these sectors arise through a WWW scenario from some field algebra and a gauge symmetry, and if so, can the latter be reconstructed from the given observables alone?

The answer in four dimensions is positive, thanks to a deep result due to Doplicher and Roberts (1990). Let us sketch the line of reasoning leading to this result in some detail, because it shows how the connection between (global) gauge symmetry on the one hand and spacetime geometry on the other hand emerges through the principle of causality (locality) of relativistic quantum field theory, and because it makes apparent what is different in low-dimensional spacetime.

The analysis is based on the general structure theory of superselection sectors due to Doplicher, Haag, and Roberts (DHR, 1971). The latter starts with a selection criterion invoking the concept of a localized charge: a superselection sector which by measurements within the causal complement of some spacetime region O cannot be distinguished from the vacuum sector. The heuristic idea is, of course, that the sector is obtained from the vacuum sector by placing some charge in the region O (e.g., by the application of a localized charged field operator to the vacuum vector).

It has been shown (Buchholz and Fredenhagen 1982) that positive-energy representations of

massive theories always satisfy this selection criterion with a localization region O of the form of a narrow cone extending in spacelike direction. (In massless theories with long-range interactions, such as QED, the situation is more complicated because the charge creates an electric field whose flux at infinity does not vanish (Gauss' law) and is not Lorentz invariant.) DHR assume that the localization region is even compact, and can be chosen arbitrarily within the unitary equivalence class of the representation.

Exploiting a strong version of locality (Haag duality) for the vacuum representation of the observables, DHR proceed to define an associative composition (or fusion) law for positive-energy representations. This law is commutative only up to unitary equivalence. The crucial point is that the unitary intertwiner establishing this equivalence (the statistics operator) can be chosen in a unique way provided any pair of spacelike disconnected localization regions can be continuously deformed into any other such pair.

This point marks the separation between high and low dimensions. In two dimensions, in each pair of spacelike disconnected regions, one region is to the left of the other, thus distinguishing the pair (O_1, O_2) from (O_2, O_1) . Consequently, they cannot be deformed into each other, and there arise two statistics operators. The same holds in three dimensions when the localization regions are spacelike cones, and O_1, O_2 are taken within (the causal complement of) some larger spacelike cone. If the spacetime dimension is at least 4, or if in three dimensions the localization regions are compact, then the statistics operator is unique and, as a consequence, coincides with its inverse.

The (non-)uniqueness of the statistics operator has far-reaching consequences concerning our original question about the underlying gauge symmetry. Namely, the DHR analysis proceeds to show that the set of positive-energy representations equipped with the composition law, and the linear spaces of intertwiners between different representations, together form the mathematical structure of a C^* -tensor category. The statistics operators which are distinguished intertwiners give additional structure to this category: this structure is called a (permutation) symmetry if the statistics operators coincide with their inverse, and it is called a braiding otherwise. (It gives rise to a representation of the permutation group or the braid group, respectively.) In other words, the spacetime topology, through the intervention of the uniqueness of the statistics operator, causes the tensor category to be symmetric in high dimensions, and braided in low dimensions.

At a more elementary level, one may think of statistics operators as reflecting commutation relations between the searched-for charged fields. Making an ansatz for the commutation relations at spacelike separation, essentially the same topological argument as before implies, together with Poincaré invariance, that the coefficients appearing in this relation should form a representation of the permutation group, or of the braid group, respectively. The DHR approach, however, is entirely intrinsic, avoiding any *a priori* assumption of charged fields.

The duality theorem due to Doplicher and Roberts (1990) now states that every symmetric C^* -tensor category (with some further qualifications valid in the DHR setting) is isomorphic to the category of unitary representations of a compact group, in which the composition law is the tensor product and the (permutation) symmetry is the natural one. Moreover, the category uniquely determines the group, and by a crossed product construction (an action of the category on the algebra A) one reconstructs a field algebra F such that [5] holds. If fermionic sectors are present, then there is some arbitrariness in the commutation relations among the corresponding fermionic fields, which can be exploited to produce the normal commutation relations (fermionic fields anticommute among each other, and bosonic fields commute with any field at spacelike separation). This fixes the field algebra F up to unitary equivalence. The conclusion is that the WWW scenario is the most general in four dimensions (apart from the reservations due to long-range forces, see above).

Generalized Symmetries in Low Dimensions

In view of the success of this program in four dimensions and the advantage of the WWW scenario for model building, the obvious challenge is to search for an analogous understanding of superselection sectors (charges) in low dimensions in terms of an algebra of charged fields and a gauge symmetry distinguishing the observables. This gauge symmetry cannot, in general, be a group for several reasons:

- As stated before, the tensor category of superselection sectors possesses only a braiding, rather than a (permutation) symmetry, hence the duality theorem fails.
- One can associate a (statistical) dimension d_π to each superselection sector $[\pi]$ which is multiplicative under the composition law (fusion), and additive under direct sums. In a symmetric

category, the dimensions are necessarily positive integers. Indeed, in the WWW scenario, they coincide with the naive dimension of the associated representation of the gauge group. But in the low-dimensional models, the dimensions turn out to be nonintegers in general.

- Moore and Seiberg (1988) have axiomatized the superselection structure of chiral and two-dimensional conformal field theories in terms of a system of recoupling and braiding coefficients controlling the fusion of sectors and its noncommutativity. (In fact, this system is basically equivalent to the DHR category.) For models such as $SU(2)$ current algebras at level k , these coefficients turn out to coincide with the recoupling and braiding coefficients one can associate with a quantum group deformation (Drinfel'd 1986) of $SU(2)$ with deformation parameter $q = -\exp i\pi/k$. Representations of quantum groups (quasitriangular Hopf algebras, see Hopf Algebras and q -Deformation Quantum Groups) have a tensor product defined in terms of a noncocommutative coproduct. Moreover, they possess a quantum dimension which is a q -deformation of an integer. The quantum dimensions precisely match the statistical dimensions of the superselection sectors. All this strongly suggests that quantum groups appear as generalized symmetries in two dimensions, at least in a large class of models.

A natural testing ground for the search for appropriate generalized symmetry concepts in low dimensions is the abundance of models in chiral and two-dimensional conformal QFT (see Two-Dimensional Models). As mentioned before, conformal symmetry in two dimensions has far-reaching consequences, especially the existence of chiral quantum fields which are defined on a one-dimensional light ray. As a null direction in the two-dimensional spacetime, this ray unites both the spacelike property of carrying a causal structure, and the timelike property that the generator of translations has positive spectrum (energy). These two features together with Möbius covariance are so powerful that they allow for the exact construction of large classes of models. The most elementary ones (minimal models) are completely described by the chiral stress-energy density field, that is, the local generator of the conformal symmetry. Other models also contain currents which are the local generators of internal symmetries. These models exhibit many nontrivial superselection structures, which illustrate the wide range of possible deviations from higher-dimensional QFT, and at the same time exhibit possible

approaches to appropriate symmetry concepts in low dimensions.

Attempts to classify the possible algebraic structures of generalized internal symmetries in a model-independent setting start from the idea that the representation category of the internal symmetries of a given model should be equivalent to the tensor category of its superselection sectors. Several algebraic structures have been proposed as candidates, complying with this idea. They all assume specific modifications or deformations of eqns [1]–[5] above, highly constrained by self-consistency. Among these proposals are:

- quantum groups (see e.g., Fröhlich and Kerler 1993),
- weak quasiquantum groups (Mack and Schomerus 1992) and rational Hopf algebras (Fuchs *et al.* 1994),
- weak C^* Hopf algebras (Rehren 1997, Böhm and Szlachányi 1996) or quantum groupoids (Nikshych and Vainerman 1998), and
- braided groups (Majid 1991).

In several cases, the respective “symmetry algebra” can be reconstructed from the tensor category of superselection sectors, and a field algebra with linear transformation behavior can be constructed which contains the observables as invariant elements as in [5]. However, the situation is unsatisfactory for various reasons. First, the class of QFT models for which these constructions have been performed is quite restricted (most constructions work only for rational models, i.e., models with a finite set of charges); second, the reconstructed symmetry algebra is not unique and finally, the constructed field algebras have features which diverge significantly from the WWW scenario. For example, it is not always warranted that the quantum symmetries are consistent with the $*$ -structure, indispensable for Hilbert space positivity (a necessary prerequisite for the probability interpretation of quantum theory). Moreover, typically there are global gauge transformations which are implemented by localized field operators, thus exhibiting a mixing of local and global concepts. It also happens that this holds for elements in the center of the symmetry algebra, which implies that the field algebra is not local relative to its gauge invariant elements, that is, the charged fields do not commute with the gauge-invariant elements at spacelike separation. In other constructions, the field algebra is not associative, or there are no finite field multiplets.

Historically, the first candidate for a “symmetry algebra” compatible with braid group statistics has

been the structure of a quantum group, as mentioned above. However, in physically interesting models, the quantum group is not semisimple and thus has too many (namely, indecomposable) representations. Solutions to this problem have been:

1. A BRS approach in an indefinite-metric framework (Hadjiivanov *et al.* 1991),
2. “Truncation,” that is, discarding the “unphysical” representations. Fröhlich and Kerler (1993) have done this consistently in a categorical framework. In fact, they have given a complete classification of the possible braided tensor categories generated by a single irreducible object with statistical dimension d satisfying $1 < d < 2$, in terms of categories constructed from the “truncated” representations of $U_q(\mathfrak{sl}_2)$. Truncation can also be performed by dividing the quantum group itself through the ideal which is annihilated by all “physical” representations, leading to a weak quasiquantum group (Mack and Schomerus 1992).
3. Relaxing the axioms, thus admitting the more general structures mentioned above.

All the above approaches assume a given generalized symmetry concept and show to what extent field algebras complying with it can be constructed. They thus concern nonobservable objects, and it is no contradiction if different symmetry concepts can be associated with the same observable data.

A more radical concept of global gauge symmetry, applicable to the low-dimensional case, has been developed by Longo and Rehren (1995). Its point of departure is the notion of a conditional expectation, which has the same abstract properties as a group average. In the WWW scenario, the Haar measure of the compact gauge group defines an average

$$\mu : F \ni \phi \mapsto \int d\mu(g) \alpha_g(\phi) \in A \quad [6]$$

which is a positive linear map respecting the localization, and the observables are invariant, $\mu(a) = a$. In fact, the observables are exactly the image of this map, that is, [5] is equivalently formulated, but without reference to the group transformations, as

$$A(O) = \mu(F(O)) \quad [7]$$

Turning to the observables A of a quantum field theory in low dimensions, one looks for a quantum field theory F , containing A and equipped with a conditional expectation μ such that [7] holds, and which preserves the vacuum state. F may not satisfy local commutativity, but it should be local relative

to the observables in the sense mentioned before. In rational chiral CFT, such extensions can be classified (and indeed constructed) in terms of the superselection category of A , giving direct access to the decomposition of the vacuum Hilbert space of F into superselection sectors of A . The advantage here is that no problems with Hilbert space structure can arise (because the approach is entirely in terms of operator algebras); a drawback is that in general F is not unique, and nonvacuum representations of F also have to be considered in order to generate all sectors of A .

The method can be used to classify and construct both nonlocal chiral extensions as candidates for sector-generating field algebras for a theory A of chiral observables, and local two-dimensional quantum field theories containing two given chiral subtheories, that is, observable algebras of two-dimensional models (Kawahigashi and Longo 2004). The chiral sector structure of the latter models is described by a “modular invariant.” In many cases, this means that their thermal partition functions are invariant under the group $\mathrm{PSL}(2, \mathbb{Z})$ of modular transformations of the temperature (see below).

At this point, another link between spacetime and internal symmetries may be noted. The modular theory of von Neumann algebras (see Tomita–Takesaki Modular Theory) associates a one-parameter group of automorphisms (called the “modular group”) with a state and an algebra “in standard position.” In quantum field theory, for the vacuum state and an algebra of observables localized in certain wedge regions of Minkowski spacetime, this group can be identified with a boost subgroup of the Lorentz group (Bisognano and Wichmann 1975). Similarly, in chiral CFT on the circle, the modular group associated with the observables in an interval and the vacuum coincides with a subgroup of the Möbius group. For nonlocal theories, there may be an obstruction, however. On the other hand, if a subalgebra is stable under the modular group of some algebra, then there is a conditional expectation from the larger algebra onto the smaller algebra. Combining these general theorems, the Möbius covariance of the inclusions $A(O) \subset F(O)$ implies the existence of a conditional expectation, that is, the above generalization of the average over the internal symmetry. Moreover, assuming a generalized notion of compactness (“finite index”) for the generalized internal symmetry, the Bisognano–Wichmann property holds also for nonlocal theories (Longo and Rehren 2004).

Of course, there is also a WWW scenario in chiral theories, that is, one may restrict a local theory to its invariants under some group of internal gauge

symmetries (“orbifold models”). It then happens that the invariants not only have the expected superselection sectors in correspondence with the representations of the gauge group, but in addition “twisted” sectors appear which, together with the former, constitute a “quantum double” structure. The twisted sectors arise by restriction of solitonic sectors of the original theory, which are in one-to-one correspondence with the elements of the gauge group (Müger 2005). Solitonic sectors are localizable with respect to two different vacua, and do not admit an unrestricted composition law.

Special Issues

A particularly simple situation is the case of anyons, that is, when all sectors have statistical dimension 1. Then the sectors form an abelian group \hat{G} under fusion, and one can construct a WWW scenario with global gauge group G the dual of \hat{G} . The ensuing quantum fields satisfy generalized commutation relations at spacelike separation, given by an abelian representation of the braid group, where the coefficients can be arbitrary complex phases (responsible for the name “anyons”). However, it is known that there can arise an obstruction, which enforces the “local” global gauge transformations (mentioned before) to be present. In this case, the gauge symmetry can also be described by a quasiquantum group. It is noteworthy that free anyon fields have been constructed in two-dimensional spacetime, while in three dimensions there can be no (cone-) localized massive anyon fields which are free in the sense that they generate only single-particle states from the vacuum (Mund 1998).

The charge structure of massive quantum field theories in two dimensions is very different both from that encountered in conformal quantum field theories, and from the charge structure in high dimensions. It has been observed long ago that, in contrast to four dimensions, the strong locality property (Haag duality) which is necessary to set up the DHR analysis of superselection sectors, fails for the algebra of invariants under an internal gauge group in two dimensions. This algebraic feature can be traced back to the fact that the causal complement of a point is disconnected in two dimensions, or, in physical terms, that “a charge cannot be transported around a detector” without passing through its region of causal dependence. Müger (1998) has shown that any algebra of observables which satisfies Haag duality, cannot possess any nontrivial DHR superselection sectors at all, and that the only sectors which can exist are solitonic

sectors. This general result nicely complies with the experience with integrable models, as mentioned before.

There are also some results giving interesting insight, which can be obtained intrinsically in terms of the observables. One of them concerns “central” observables (generalized Casimir operators).

Casimir operators in the WWW scenario are functions of the generators of the internal symmetry which usually are integrals over densities belonging to the field algebra F (Noether’s theorem). Since they also commute with the generators, they can be approximated by local observables, and are therefore defined in each representation of the latter. By Schur’s lemma, they are multiples of the identity in each irreducible sector. Since the eigenvalues of Casimir operators distinguish the representations of the gauge group, they also distinguish the sectors.

In chiral CFT extended to the circle (see above), one can find global “charge measuring operators” C_i , one for each sector π_i , in the center of the observable algebra (Fredenhagen *et al.* 1992) which have similar properties. They arise as a consequence of an algebraic obstruction to define the charged sectors on the circle, related to a nontrivial effect if a charge is “transported once around the circle,” and form an operator representation of the fusion rules within the global algebra of observables. Under rather natural conditions clarified by Kawahigashi, Longo, and Müger (2001), the matrix of eigenvalues $\pi_j(C_i)$ is nondegenerate, that is, the generalized Casimir operators completely distinguish the superselection sectors. In this case, the superselection category is a modular category (see Braided and Modular Tensor Categories): the matrix with entries $d_{\pi_i} \pi_j(C_i)$ and the diagonal matrix with entries $\pi_j(U)$ (where U is the Möbius rotation by 2π) are multiples of the generators S and T of the “modular group” $\text{PSL}(2, \mathbb{Z})$, in a matrix representation labeled by the superselection sectors of the chiral observables. The physical significance of this matrix representation is that it relates thermal expectation values for different values of the temperature (Cardy 1986, Kac and Peterson 1984, Verlinde 1988).

These examples, together with the failure of the Coleman–Mandula theorem, may illustrate the intricate relations among spacetime geometry, covariance, and internal symmetry (charge structure) in low dimensions. In relativistic quantum field theory, the link is provided by the principle of locality, which “turns geometry into algebra.”

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Braided and Modular Tensor Categories; Hopf Algebras and q -Deformation

Quantum Groups; Integrability and Quantum Field Theory; Quantum Field Theory: A Brief Introduction; Quantum Fields with Topological Defects; Symmetries and Conservation Laws; Symmetries in Quantum Field Theory: Algebraic Aspects; Symmetry Breaking in Field Theory; Tomita–Takesaki Modular Theory; Two-Dimensional Conformal Field Theory and Vertex Operator Algebras; Two-Dimensional Models.

Further Reading

- Böhm G and Szlachányi K (1996) A coassociative C^* -quantum group with non-integral dimensions. *Letters in Mathematical Physics* 35: 437–448.
- Coleman S and Mandula J (1967) All possible symmetries of the S -matrix. *Physical Review* 159: 1251–1256.
- Doplicher S and Roberts JE (1990) Why there is a field algebra with a compact gauge group describing the superselection structure in particle physics. *Communications in Mathematical Physics* 131: 51–107.
- Fredenhagen K, Rehren K-H, and Schroer B (1992) Superselection sectors with braid group statistics and exchange algebras II: geometric aspects and conformal covariance. *Reviews in Mathematical Physics* S11: 113–157.
- Fröhlich J and Kerler T (1993) *Quantum Groups, Quantum Categories, and Quantum Field Theory*. Lecture Notes in Mathematics, vol. 1542. Berlin: Springer.
- Fuchs J, Ganchev A, and Vecsernyés P (1994) On the quantum symmetry of rational field theories. *Theoretical Mathematical Physics* 98: 266–276.
- Hadjiivanov LK, Paunov RR, and Todorov IT (1991) Quantum group extended chiral p -models. *Nuclear Physics B* 356: 387–438.
- Longo R and Rehren K-H (1995) Nets of subfactors. *Rev. Mathematical Physics* 7: 567–597.
- Mack G and Schomerus V (1992) Quasi Hopf quantum symmetry in quantum theory. *Nuclear Physics B* 370: 185–230.
- Majid S (1991) Braided groups and algebraic quantum field theories. *Letters in Mathematical Physics* 22: 167–175.
- Moore G and Seiberg N (1989) Classical and quantum conformal field theory. *Communications in Mathematical Physics* 123: 177–254.
- Müger M (1998) Superselection structure of massive quantum field theories in $1+1$ dimensions. *Reviews in Mathematical Physics* 10: 1147–1170.
- Mund J (1998) No-go theorem for “free” relativistic anyons in $d=2+1$. *Letters in Mathematical Physics* 43: 319–328.
- Rehren K-H (1997) Weak C^* Hopf symmetry. In: Doebner H-D and Dobrev VK (eds.) *Group Theoretical Methods in Physics*, pp. 62–69. (q-alg/9611007). Sofia: Heron Press.
- Schomerus V (1995) Construction of field algebras with quantum symmetries from local observables. *Communications in Mathematical Physics* 169: 193–236.
- Wick GC, Wightman AS, and Wigner EP (1952) The intrinsic parity of elementary particles. *Physical Review* 88: 101–105.

Symmetries in Quantum Field Theory: Algebraic Aspects

J E Roberts, Università di Roma “Tor Vergata,”
Rome, Italy

© 2006 Elsevier Ltd. All rights reserved.

Introduction

This article treats the most important results and concepts relating to symmetry and conservation laws in quantum field theory. It includes such results as Wigner’s theorem, Goldstone’s theorem, the Bisognano–Wichmann theorem, the quantum Noether theorem, and the theorem on the existence of gauge groups and a field net. It is written within the framework of algebraic quantum field theory, this being the simplest setting capable of expressing all these concepts and results.

Symmetries come in many guises. They are to a physical system what automorphisms are to a mathematical theory. In fact, when a physical system is described in mathematical terms, its symmetries correspond to the automorphisms of the mathematical structure and in particular form a group, its symmetry group. The reader should bear in mind this simple picture throughout its diverse

variations. Readers unfamiliar with the mathematical terminology should consult the appendix.

Elementary Quantum Mechanics

Before turning to quantum field theory, let us comment on symmetries in elementary quantum mechanics. These systems have the density matrices, that is, positive operators of trace 1, on an infinite-dimensional separable Hilbert space as states, the self-adjoint operators as observables. The expectation value of the bounded observable A in the state determined by ρ is given by $\text{tr } \rho A$. Having specified the mathematical structure, the notion of symmetry follows. With a suggestive notation, it is a pair of mappings $A \mapsto \alpha A, \rho \mapsto \rho \alpha^{-1}$ such that

$$\text{tr } \rho \alpha^{-1} \alpha A = \text{tr } \rho A$$

for all observables A and states ρ .

If we take ρ and A to be the projections onto $C\phi$ and $C\psi$ for unit vectors ϕ and ψ , then the above condition corresponds to the conservation of

transition probabilities $|(\phi, \psi)|^2$. This formed the starting point for Wigner's analysis, who concluded:

Theorem *Every symmetry is of the form $A \mapsto UAU^{-1}$ and $\rho \mapsto U\rho U^{-1}$, where U is a unitary or antiunitary operator.*

As could have been foreseen from the outset, this simple result in no way distinguishes one elementary quantum-mechanical system from another. A more useful notion of symmetry results if the Hamiltonian is reckoned as part of the information describing the system and, therefore, has to be left invariant by a symmetry. The operator U above must therefore satisfy the condition $UHU^{-1} = H$ and it commutes with the Hamiltonian. As the Hamiltonian is the generator of time translations, U is a constant of motion. This is the genesis of the relation between symmetries and conservation laws.

Quantum Field Theories

The simplest types of quantum field theories can be described by von Neumann algebras $\mathfrak{A}(\mathcal{O})$ depending on double cones \mathcal{O} and subject to

$$\mathcal{O}_1 \subset \mathcal{O}_2 \Rightarrow \mathfrak{A}(\mathcal{O}_1) \subset \mathfrak{A}(\mathcal{O}_2)$$

a structure referred to as the net of observables.

An alternative approach would be to use the Wightman formalism. This would need a discussion of pointlike fields and the domains of definition of unbounded operators, thus complicating a general exposition of symmetry.

Comparing this description of a quantum field theory with that of an elementary quantum-mechanical system, the net clearly substitutes observables but nothing has yet been said about states. Since the set of double cones is directed under inclusion, the union of the $\mathfrak{A}(\mathcal{O})$ is a $*$ -algebra \mathfrak{A} and a state of our system is a state on this algebra.

Most states are of no physical relevance. A characterization of the states of physical relevance, even say to elementary particle physics, is not known although some progress has been made.

The net structure is the hallmark of a field theory and allows us to distinguish two important classes of symmetries. An internal symmetry α satisfies the condition

$$\alpha(\mathfrak{A}(\mathcal{O})) = \mathfrak{A}(\mathcal{O})$$

for all double cones \mathcal{O} . By contrast, a spacetime symmetry is an automorphism α_L implementing a Poincaré transformation L and hence satisfying the condition

$$\alpha_L(\mathfrak{A}(\mathcal{O})) = \mathfrak{A}(L\mathcal{O})$$

for every double cone \mathcal{O} . It is usually the case that internal symmetries commute with spacetime symmetries.

The state of prime relevance to elementary particle physics is the vacuum state ω_0 . The corresponding Gelfand–Naimark–Segal (GNS) representation π_0 is called the vacuum representation. Now the vacuum state of a quantum field theory is typically unique and as such invariant under a symmetry of the system $\omega_0\alpha^{-1} = \omega_0$.

Spacetime Symmetries

Since the vacuum state is invariant, we have a unitary representation of the Poincaré group implementing the spacetime symmetries in the vacuum representation. To illustrate the role of representations up to a factor, we take instead the GNS representation of a pure state corresponding to a particle of half-integral spin. Here we need a unitary representation of the covering group of the Poincaré group, inhomogeneous $SL(2, \mathbb{C})$ to implement the symmetries. The situation for the subgroup of rotations is the same.

The most important property of these representations is positivity of the energy. More precisely, in a representation of relevance to elementary particle physics such as the vacuum representation, the generator P^0 of time translations is a positive operator $P^0 \geq 0$. Expressed in a frame-independent way, the spectrum of spacetime translations is contained in the closed forward light cone. It is one of the basic principles to be exploited in applying quantum field theory to elementary particle physics. Notice that the principle is no longer valid for an equilibrium state.

A similar situation arises in conformal field theory. Here the role of double cones in Minkowski space is played by intervals on the circle and that of the Poincaré group by the Möbius group on the circle $PSL(2, \mathbb{R})$. Again, the Möbius group cannot always be unitarily implemented and conformal invariance is defined via a continuous unitary representation of its covering group. Most importantly, there is an analog of positivity of the energy. The generator of rotations of the circle is a positive operator.

A remarkable aspect of spacetime symmetries was discovered by Bisognano and Wichmann in an application of modular theory in the field-theoretical context looking not at double cones but at wedges. A wedge \mathcal{W} is a Poincaré transform of the standard wedge $x^1 > |x^0|$. They found that the modular automorphisms of $\mathfrak{A}(\mathcal{W})$ and the vacuum vector Ω_0

have a geometric significance. For the standard wedge, they got the following result.

Theorem *If the net is derived from Wightman fields, the modular operator is $e^{-2\pi K}$, where K is the generator of boosts in the 1-direction and the modular conjugation is $ZR\Theta$, where Θ is the TCP-operator, R is the rotation through π about the 1-axis, and Z is the unitary operator equal to 1 on the Bose subspace and $-i$ on the Fermi subspace.*

The modular data for $\mathfrak{A}(\mathcal{O})$ and Ω_0 also admit a geometric interpretation for the free massless scalar field.

These facts enhance our understanding of space-time symmetries. The ideas have meanwhile been applied to curved spacetime to select a state with vacuum-like properties using the principle of the geometric action of the modular conjugation.

Gauge Symmetry

Gauge symmetries do not fit into our scheme in that they act trivially on the observable algebra \mathfrak{A} . To exhibit a gauge symmetry we need a larger net \mathfrak{F} called the field net. The gauge group will be the group of automorphisms of \mathfrak{F} leaving the subnet \mathfrak{A} pointwise fixed and \mathfrak{A} the subnet of \mathfrak{F} of fixed points under G . This has the merit of indicating the mathematical framework for gauge symmetry but otherwise begs important questions. *A priori* one does not know what properties \mathfrak{F} should have nor how it should be constructed.

The right approach is to understand what intrinsic structure of \mathfrak{A} governs the existence of a nontrivial gauge group. This brings us back to the states or representations relevant to elementary particle physics. A condition for selecting some of these relevant representations is that asymptotically they be like the vacuum in spacelike directions. More precisely, π must be unitarily equivalent to the vacuum representation π_0 on the spacelike complement of every double cone.

The resulting theory of superselection sectors hinges on the property of Haag duality that, for each double cone \mathcal{O} ,

$$\mathfrak{A}(\mathcal{O}) = \mathfrak{A}(\mathcal{O}')'$$

where \mathcal{O}' denotes the spacelike complement of \mathcal{O} . It implies that every representation satisfying the selection criterion is unitarily equivalent to one of the form $\pi_0\rho$, where ρ is an endomorphism of \mathfrak{A} localized in some fixed but arbitrary double cone, that is, $\rho(A) = A$ if $A \in \mathfrak{A}(\mathcal{O}')$. The endomorphisms thus obtained are closed under composition and

hence the objects of a full tensor subcategory \mathcal{T} of the category of all endomorphisms and their intertwiners. There is a dimension function d defined on the objects of \mathcal{T} , $d(\rho) = 1, 2, \dots, \infty$. If \mathcal{T}_f denotes the full subcategory whose objects have finite dimension, then the following result holds.

Theorem *\mathcal{T}_f is equivalent to the tensor category of finite-dimensional continuous unitary representations of a canonical compact group G . There is a canonical field net \mathfrak{F} with Bose–Fermi commutation relations extending \mathfrak{A} such that G is the group of automorphisms of \mathfrak{F} leaving \mathfrak{A} pointwise fixed.*

The first step in the proof is to define and analyze the statistics of the representations in question. The statistics of an irreducible representation ρ can be classified as being para-Bose or para-Fermi of order $d(\rho)$. The second step is to show that each ρ of finite dimension has a well-defined conjugate up to equivalence. The third and most difficult step is showing that \mathcal{T}_f can be embedded in the tensor category of Hilbert spaces.

The Local Implementation of Symmetries

Gauge symmetry has its associated conservation laws in that the different sectors of the last section are labeled by conserved quantities such as baryon number, lepton number, or electric charge, generically called charges. The theory is built round the idea of creating charge and elements of the field net carry charges. But there should be a dual approach based on measuring charges. One would like to prove the existence of local conserved currents corresponding to these charges. This has not proved possible but there is a good substitute, described below, which can be regarded as a weak version of a quantum Noether theorem.

If $\mathcal{O}_1 \subset \mathcal{O}_2$ is a strict inclusion of double cones, then the theory is said to satisfy the split property if there is a type I factor \mathcal{M} such that

$$\mathfrak{A}(\mathcal{O}_1) \subset \mathcal{M} \subset \mathfrak{A}(\mathcal{O}_2)$$

where a type I factor is a von Neumann algebra isomorphic to some $B(\mathcal{H})$. In this case \mathcal{M} can be chosen in a canonical fashion and there is an isomorphism ψ called the universal localizing map of $B(\mathcal{H})$ onto \mathcal{M} , where \mathcal{H} is the underlying Hilbert space. We have $\psi(A) = A$ for $A \in \mathfrak{A}(\mathcal{O}_1)$.

Theorem *If U is an implementing representation of the internal symmetry group G , $\psi(U)$ will be a representation of G in \mathcal{M} that continues to implement the symmetry on $\mathfrak{A}(\mathcal{O}_1)$. If G is a Lie group*

then the infinitesimal generators in the representation are an analog of locally integrated current densities.

Spontaneously Broken Symmetry

The standard physical example of a spontaneously broken symmetry is magnetization. Despite the overall rotational symmetry, a magnet picks out a preferred direction as its direction of magnetization. The chosen state breaks the symmetry.

The phenomenon of spontaneously broken symmetry involves an interplay of symmetries and certain classes of states, vacuum states, ground states, or equilibrium states. If such an ω is induced by a vector cyclic and separating for a local algebra $\mathfrak{A}(\mathcal{O})$, then, as explained in the appendix, given \mathcal{O} , modular theory yields a canonical unitary representation V of the internal symmetry group G :

$$gA = V_g A V_g^*, \quad A \in \mathfrak{A}(\mathcal{O})$$

The results concern the breaking of a one-parameter group $\lambda \mapsto \alpha_\lambda$ of symmetries. More precisely, one asks whether $\omega\delta=0$ or not, where δ is the infinitesimal generator of $\lambda \mapsto \alpha_\lambda$,

$$\delta(F) = \lim_{\lambda \rightarrow 0} \lambda^{-1}(\alpha_\lambda(F) - F)$$

where norm convergence is understood and holds on a dense domain. δ , the derivation, is an infinitesimal symmetry. Goldstone first showed that the spontaneous breaking of such symmetries requires the presence of massless bosons. The following result is taken from a more modern treatment. \mathcal{O}_R here denotes the double cone whose base is the ball in $t=0$ of radius R centered on the origin and \mathcal{D} the domain of δ .

Theorem *Let δ be a derivation on a field net \mathfrak{F} in $s > 1$ spatial dimensions such that for $F \in \mathfrak{F}(\mathcal{O}_R) \cap \mathcal{D}$*

$$|\omega_0 \delta F| \leq c_{R,\varepsilon}(\|F\Omega\| + \|F^*\Omega\|) + \varepsilon\|\delta F\|$$

- (i) *If $\liminf_{R \rightarrow \infty} c_{R,\varepsilon} R^{-(s-1)/2} = 0$, then $\omega_0 \delta = 0$.*
- (ii) *If $\liminf_{R \rightarrow \infty} c_{R,\varepsilon} R^{-(s-1)/2} < \infty$, then $\omega_0 \delta \neq 0$ is only possible if the spectrum of the translations coincides with the forward light cone V_+ and the boundary $\partial V_+ \setminus \{0\}$ has non-trivial spectral measure (i.e., there are massless particles in the theory).*
- (iii) *If $c_{R,\varepsilon}$ is polynomially bounded in R , then $\omega_0 \delta \neq 0$ is only possible if the spectrum of translations coincides with V_+ but there are not necessarily any massless particles.*

Symmetries of the S-matrix

Scattering theory not only allows one to construct the multiparticle scattering states but also shows that internal symmetries and spacetime symmetries continue to act on these states and are therefore symmetries of the S -matrix. We can, however, ask what are all the symmetries of the S -matrix. An answer was provided by Coleman and Mandula, who showed that, when there is nontrivial scattering, there are no further symmetries of the S -matrix.

Appendix

In an effort to make this article more self-contained, this appendix collects together a few simple pertinent concepts and results from the theory of operator algebras. A C^* -algebra is a $*$ -algebra \mathcal{A} with a norm $\|\cdot\|$ making it into a Banach algebra and satisfying

$$\|A^*A\| = \|A\|^2$$

for every $A \in \mathcal{A}$. Any C^* -algebra can be realized as a norm closed $*$ -subalgebra of the C^* -algebra $\mathcal{B}(\mathcal{H})$ of all bounded operators on a Hilbert space \mathcal{H} . A von Neumann algebra \mathcal{R} is a C^* -algebra that is the dual space of a Banach space. This Banach space \mathcal{R}_* , the predual of \mathcal{R} , is intrinsically defined. The topology on \mathcal{R} determined by duality with \mathcal{R}_* is called the σ -topology. $\mathcal{B}(\mathcal{H})$ is a von Neumann algebra and its predual is the set of trace class operators. Any von Neumann algebra can be realized as a σ -closed unital $*$ -subalgebra of some $\mathcal{B}(\mathcal{H})$.

A state on a C^* -algebra \mathcal{A} is a positive linear functional ω of norm 1. If \mathcal{A} has a unit I the normalization condition can be expressed as $\omega(I) = 1$. Of fundamental importance is the relation between representations and states. A representation of \mathcal{A} on a Hilbert space \mathcal{H} is just a structure-preserving mapping or morphism of \mathcal{A} into $\mathcal{B}(\mathcal{H})$. For simplicity, we suppose that \mathcal{A} has a unit. Given a state ω , there is an associated representation π_ω defined by a vector Ω such that $\pi_\omega(\mathcal{A})\Omega$ is dense in the Hilbert space in question, that is, it is a cyclic vector for the representation and

$$\omega(A) = (\Omega, \pi_\omega(A)\Omega), \quad A \in \mathcal{A}$$

that is, the cyclic vector implements the given state. This is referred to as the GNS construction. Given any two such representations, there is a unique unitary operator mapping the one cyclic vector onto the other and realizing the equivalence of the representations.

A state of a von Neumann algebra is said to be normal if it is continuous in the σ -topology. If ω is normal, then $\pi_\omega(\mathcal{R})$ is σ -closed.

An inclusion of unital von Neumann algebras has the split property if there is an intermediate type I factor, that is, if it has the form $\mathcal{R}_1 \subset \mathcal{B}(\mathcal{H}) \subset \mathcal{R}_2$.

The following elementary observation is often used in treating symmetries. If α is an automorphism of \mathcal{A} with $\omega\alpha^{-1} = \omega$, there is a unique unitary operator leaving the cyclic vector Ω invariant and inducing α in the representation π_ω . In other words, $U\Omega = \Omega$ and

$$U\pi_\omega(A)U^{-1} = \pi_\omega(\alpha A)$$

If we apply the above lemma to a group G of symmetries leaving a state invariant, it yields a group $U(g)$ of unitaries satisfying the condition

$$U(gh) = U(g)U(h), \quad g, h \in G$$

since $U(g)$ is uniquely defined by the above conditions.

When there is no invariant state, the situation is more complicated. Suppose there is a group G of symmetries and a representation π of \mathfrak{A} where each g is unitarily implemented. Thus, there is a unitary $U(g)$ with

$$U(g)\pi(A)U(g)^{-1} = \pi(gA), \quad A \in \mathfrak{A}$$

All we can now conclude is that

$$U(gh) = Z(g, h)U(g)U(h)$$

where $Z(g, h)$ is a unitary in \mathfrak{A}' , the commutant of \mathfrak{A} , satisfying the 2-cocycle identity

$$Z(gh, k)Z(g, h) = Z(g, hk)^g Z(h, k)$$

where $^g X = U(g)XU(g)^{-1}$. U is said to be a representation up to a factor. It can be chosen to be a representation if the cocycle Z is a coboundary, that is, if there is a unitary $Y(g)$ in \mathfrak{A}' such that

$$Y(g)^g Y(h) = Y(gh)Z(g, h)$$

In general, little is known about solving problems of this kind, but there are a number of results when π is irreducible and the unitary group of its commutant reduces to the circle.

We turn now to consider the modular theory of von Neumann algebras. A vector Ω is said to be separating for a von Neumann algebra \mathcal{R} if $A\Omega = 0$ and $A \in \mathcal{R}$ implies $A = 0$. If Ω is both cyclic and separating, there is a uniquely determined closed antilinear involution S with $SA\Omega = A^*\Omega$ for $A \in \mathcal{R}$. If $S = J\Delta^{1/2}$ is the polar decomposition of S , then the unitary operators Δ^{it} induce automorphisms δ^{it} of \mathcal{R}

and $J\mathcal{R}J = \mathcal{R}'$. J is called the modular conjugation, Δ the modular operator, and δ^{it} the modular automorphisms. The closure of $\{\Delta^{1/4}A\Omega : A \in \mathcal{R}, A \geq 0\}$ is a cone, called the natural cone. Every normal state of \mathcal{R} is implemented by a unique vector in the natural cone. If α is an automorphism of \mathcal{R} , there is therefore a unique vector Ω_α in the natural cone such that, for every $A \in \mathcal{R}$,

$$(\Omega, \alpha^{-1}(A)\Omega) = (\Omega_\alpha, A\Omega_\alpha)$$

There is now a canonical unitary operator V_α defined by

$$V_\alpha A\Omega = \alpha(A)\Omega_\alpha$$

V_α maps the natural cone into itself and $\alpha \mapsto V_\alpha$ is an implementing representation of the group of automorphisms of \mathcal{R} . Under these circumstances, we do not have to deal with representations up to a factor.

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Boundary Conformal Field Theory; Current Algebra; Quantum Fields with Topological Defects; Supergravity; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Two-Dimensional Models.

Further Reading

- Bisognano JJ and Wichmann EH (1976) On the duality condition for quantum fields. *Journal of Mathematical Physics* 17: 303–321.
- Borchers HJ and Buchholz D (1985) The energy-momentum spectrum in local field theories with broken Lorentz symmetry. *Communications in Mathematical Physics* 97: 169–185.
- Buchholz D (1982) The physical state space of quantum electrodynamics. *Communications in Mathematical Physics* 85: 49–71.
- Buchholz D, Doplicher S, and Longo R (1986) On Noether's theorem in quantum field theory. *Annals of Physics* 170: 1–17.
- Buchholz D, Doplicher S, Longo R, and Roberts JE (1992) A new look at Goldstone's theorem. *Reviews in Mathematical Physics* (Special Issue): 49–83.
- Buchholz D, Dreyer O, Florig M, and Summers SJ (2000) Geometric modular action and spacetime symmetry groups. *Reviews in Mathematical Physics* 12: 475–560.
- Coleman S and Mandula J (1967) All possible symmetries of the S-matrix. *Physical Review* 159: 1251–1256.
- Doplicher S and Roberts JE (1990) Why there is a field algebra with a compact gauge group describing the superselection structure in particle physics. *Communications in Mathematical Physics* 131: 51–107.
- Guido D and Longo R (1996) The conformal spin and statistics theorem. *Communications in Mathematical Physics* 181: 11–35.
- Lopuszański J (1991) *An Introduction to Symmetry and Supersymmetry in Quantum Field Theory*. Singapore: World Scientific.
- O'Raifeartaigh L (1965) Lorentz invariance and internal symmetry. *Physical Review* 139: B1052–B1062.

Symmetry and Symmetry Breaking in Dynamical Systems

I Melbourne, University of Surrey, Guildford, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The same symmetries may underlie diverse contexts such as phase transitions of crystals (Landau theory), fluid dynamics, and problems in biology and chemical engineering. Hence, seemingly unrelated systems may exhibit similar phenomena in regard to symmetries of patterns and transitions between patterns (spontaneous symmetry breaking). It is natural to focus attention on aspects of pattern formation that are universal or model independent – aspects depending on underlying symmetries rather than model-specific details.

The general framework is that the underlying system is governed by an evolution equation

$$\dot{x} = f(x) \quad [1]$$

with symmetry group Γ . To avoid technicalities, we assume that [1] is an ordinary differential equation (ODE), the vector field $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is as smooth as desired, and Γ is a compact Lie group acting linearly on \mathbb{R}^n . An inner product may be chosen so that Γ acts orthogonally. The vector field in [1] is Γ -equivariant if

$$f(\gamma x) = \gamma f(x) \quad \text{for all } x \in \mathbb{R}^n, \gamma \in \Gamma \quad [2]$$

Equivalently, if $x(t)$ is a solution and $\gamma \in \Gamma$, then $\gamma x(t)$ is a solution.

In this article, we are interested in the dynamics to be expected for equivariant vector fields, and transitions that arise as parameters are varied. The symmetry group Γ is taken as given, whereas f is a general Γ -equivariant vector field. (Other features such as energy conservation or time reversibility must be built into the general setup, but are excluded in this article.)

Isotropy Subgroups and Commuting Linear Maps

Let Γ be a compact Lie group acting linearly on \mathbb{R}^n . The isotropy subgroup of $x \in \mathbb{R}^n$ is defined to be

$$\Sigma_x = \{\gamma \in \Gamma: \gamma x = x\}$$

Note that $\Sigma_{\gamma x} = \gamma \Sigma_x \gamma^{-1}$ for all $x \in \mathbb{R}^n, \gamma \in \Gamma$.

Given an isotropy subgroup $\Sigma \subset \Gamma$, define the fixed-point subspace

$$\text{Fix } \Sigma = \{y \in \mathbb{R}^n: \sigma y = y \text{ for all } \sigma \in \Sigma\}$$

If $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a Γ -equivariant vector field, then $f(\text{Fix } \Sigma) \subset \text{Fix } \Sigma$ for each isotropy subgroup Σ . Hence $\text{Fix } \Sigma$ is flow invariant.

The normalizer $N(\Sigma) = \{\gamma \in \Gamma: \gamma \Sigma \gamma^{-1} = \Sigma\}$ is the largest subgroup of Γ that acts on $\text{Fix } \Sigma$, and $f|_{\text{Fix } \Sigma}$ is $(N(\Sigma)/\Sigma)$ -equivariant.

An isotropy subgroup Σ is axial if $\dim \text{Fix } \Sigma = 1$, and then $N(\Sigma)/\Sigma \cong \mathbb{Z}_2$ or 1 . More generally, Σ is maximal if there are no isotropy subgroups T with $\Sigma \subset T \subset \Gamma$ other than $T = \Sigma$ and $T = \Gamma$. Then $N(\Sigma)/\Sigma$ acts fixed-point freely on $\text{Fix } \Sigma$ and the connected component of the identity $(N(\Sigma)/\Sigma)^0 \cong 1, \text{SO}(2)$ or $\text{SU}(2)$. Correspondingly Σ is called real, complex, or quaternionic. In the complex case $\dim \text{Fix } \Sigma$ is even; in the quaternionic case $\dim \text{Fix } \Sigma \equiv 0 \pmod{4}$.

The dihedral group $\Gamma = D_m$ of order m is the symmetry group of the regular m -gon, $m \geq 3$. Its standard action on \mathbb{R}^2 is generated by

$$\rho = \begin{pmatrix} \cos 2\pi/m & -\sin 2\pi/m \\ \sin 2\pi/m & \cos 2\pi/m \end{pmatrix}$$

$$\kappa = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

For m even, the isotropy subgroups up to conjugacy are

$$D_m, \quad \mathbb{Z}_2(\kappa), \quad \mathbb{Z}_2(\rho\kappa), \quad 1$$

where $\mathbb{Z}_j(g)$ denotes the cyclic group of order j generated by g . The maximal isotropy subgroups $\Sigma = \mathbb{Z}_2(\kappa), \mathbb{Z}_2(\rho\kappa)$ are axial with $N(\Sigma)/\Sigma \cong \mathbb{Z}_2$. For m odd, $\mathbb{Z}_2(\rho\kappa)$ is conjugate to $\mathbb{Z}_2(\kappa)$ leaving three conjugacy classes of isotropy subgroups, and $\Sigma = \mathbb{Z}_2(\kappa)$ is axial with $N(\Sigma)/\Sigma = 1$.

The space of commuting linear maps

$$\text{Hom}_\Gamma(\mathbb{R}^n) = \{L: \mathbb{R}^n \rightarrow \mathbb{R}^n \text{ linear: } L\gamma = \gamma L \text{ for all } \gamma \in \Gamma\}$$

is completely described representation-theoretically. Recall that Γ acts irreducibly on \mathbb{R}^n if the only Γ -invariant subspaces of \mathbb{R}^n are \mathbb{R}^n and $\{0\}$. Then $\text{Hom}_\Gamma(\mathbb{R}^n)$ is a real division ring (skew field) $\mathcal{D} \cong \mathbb{R}, \mathbb{C}$ or \mathbb{H} . The representation is called absolutely irreducible when $\mathcal{D} = \mathbb{R}$ and nonabsolutely irreducible when $\mathcal{D} = \mathbb{C}$ or \mathbb{H} .

If the action of Γ is not irreducible, write $\mathbb{R}^n = V_1 \oplus \cdots \oplus V_k$ (nonuniquely) as a sum of irreducible subspaces. Summing together irreducible subspaces that are isomorphic to form isotypic components W gives the (unique) isotypic decomposition $\mathbb{R}^n = W_1 \oplus \cdots \oplus W_\ell$. If $L \in \text{Hom}_\Gamma(\mathbb{R}^n)$, then $L(W_j) \subset W_j$ for each j , hence $\text{Hom}_\Gamma(\mathbb{R}^n) = \text{Hom}_\Gamma(W_1) \oplus \cdots \oplus \text{Hom}_\Gamma(W_\ell)$. Each W_j consists of k_j isomorphic copies of an irreducible representation with division ring \mathcal{D}_j . Let $M_k(\mathcal{D})$ denote the space of $k \times k$ matrices with entries in \mathcal{D} . Then

$$\text{Hom}_\Gamma(\mathbb{R}^n) \cong M_{k_1}(\mathcal{D}_1) \oplus \cdots \oplus M_{k_\ell}(\mathcal{D}_\ell) \quad [3]$$

Spectral properties of commuting linear maps can be recovered from the decomposition [3], paying due attention to multiplicity and complex conjugates of eigenvalues.

Equivariant Dynamics

The dynamics of equivariant systems includes (relative) equilibria and periodic solutions, robust heteroclinic cycles/networks, and symmetric chaotic attractors.

Equilibria

Consider the ODE [1] with Γ -equivariant vector field f satisfying [2]. If $x(t) \equiv x_0$ is an equilibrium, $f(x_0) = 0$, then there is a group orbit Γx_0 of equilibria.

Let $\Sigma = \Sigma_{x_0}$ be the isotropy subgroup of x_0 . If $\dim \Sigma = \dim \Gamma$, then generically (for an open dense set of Γ -equivariant vector fields), the eigenvalues of $(df)_{x_0}$ have nonzero real part, hence x_0 is hyperbolic. If the eigenvalues all have negative real part, then x_0 is asymptotically stable. If at least one eigenvalue has positive real part, then x_0 is unstable. Hyperbolic equilibria are isolated and persist under perturbations of f ; the perturbed equilibria continue to have isotropy Σ . Since $(df)_{x_0} \in \text{Hom}_\Sigma(\mathbb{R}^n)$, decomposition [3] for the action of Σ on \mathbb{R}^n facilitates stability computations for x_0 .

If $\dim \Sigma < \dim \Gamma$, then Γx_0 is a continuous group orbit of equilibria. Generically, $\dim \ker(df)_{x_0} = \dim \Gamma - \dim \Sigma$ and $\ker(df)_0 = \{\xi x_0 : \xi \in L\Gamma\}$, where $L\Gamma$ is the Lie algebra of Γ . The remaining $k = n - \dim \Gamma + \dim \Sigma$ eigenvalues generically have nonzero real part so Γx_0 is normally hyperbolic. If all k eigenvalues have nonzero real part, then Γx_0 is asymptotically stable. If at least one has positive real part, then Γx_0 is unstable. When $N(\Sigma)/\Sigma$ is finite, generically x_0 is an isolated equilibrium in $\text{Fix } \Sigma$ and persists as an equilibrium with isotropy Σ under perturbation.

Relative Equilibria and Skew Products

A point $x_0 \in \mathbb{R}^n$ (or the corresponding group orbit Γx_0) is a relative equilibrium if $f(x_0) \in T_{x_0} \Gamma x_0 = L\Gamma x_0$. If x_0 has isotropy Σ , then x_0 is a relative equilibrium if $f(x_0) \in LD_\Sigma x_0$, where $D_\Sigma = (N(\Sigma)/\Sigma)^0$.

Write $f(x_0) = \xi x_0$, where $\xi \in LD_\Sigma$. The closure of the one-parameter subgroup $\exp(t\xi)$ is a maximal torus in D_Σ for almost every ξ . All maximal tori are conjugate with common dimension $d = \text{rank } D_\Sigma$. The solution $x(t) = \exp(t\xi)x_0$ is typically a d -dimensional quasiperiodic motion. "Typically" holds in both the topological and probabilistic sense and there is no phase-locking. When $d = 1$, $x(t)$ is periodic, often called a rotating wave.

Choose a Σ -invariant local cross section X to the group orbit Γx_0 at x_0 . There is a Γ -invariant neighborhood of Γx_0 that is Γ -equivariantly diffeomorphic to $(\Gamma \times X)/\Sigma$, where Σ acts freely on $\Gamma \times X$ by

$$\sigma \cdot (\gamma, x) = (\gamma\sigma^{-1}, \sigma x)$$

and Γ acts by left multiplication on the first factor. The Γ -equivariant ODE on $(\Gamma \times X)/\Sigma$ lifts to a $(\Gamma \times \Sigma)$ -equivariant skew product on $\Gamma \times X$

$$\dot{\gamma} = \gamma\xi(x), \quad \dot{x} = h(x) \quad [4]$$

where $\xi : X \rightarrow L\Gamma$, $h : X \rightarrow X$ satisfy the Σ -equivariance conditions

$$\begin{aligned} \xi(\sigma x) &= \text{Ad}_\sigma \xi(x) = \sigma \xi(x) \sigma^{-1} \\ h(\sigma x) &= \sigma h(x) \end{aligned}$$

and $h(x_0) = 0$.

Thus, dynamics near the relative equilibrium $\Gamma x_0 \subset \mathbb{R}^n$ reduces to dynamics near the ordinary equilibrium $x_0 \in X$ for the Σ -equivariant vector $h : X \rightarrow X$, coupled with Γ drifts. In particular, the stability of Γx_0 is determined by $(dh)_{x_0}$.

Periodic Solutions

A nonequilibrium solution $x(t)$ is periodic if $x(t+T) = x(t)$ for some $T > 0$. The least such T is the (absolute) period. The spatial symmetry group Δ is the isotropy subgroup of $x(t)$ for some, and hence all, $t \in \mathbb{R}$. The periodic solution $P = \{x(t) : 0 \leq t < T\}$ lies inside $\text{Fix } \Delta$. Define the spatiotemporal symmetry group $\Sigma = \{\gamma \in \Gamma : \gamma P = P\}$. Note that Δ is a normal subgroup of Σ and either $\Sigma/\Delta \cong S^1$ (P is a rotating wave) or $\Sigma/\Delta \cong \mathbb{Z}_q$ and P is called a standing wave or a discrete rotating wave. For each $\sigma \in \Sigma$, there exists $T_\sigma \in [0, T)$ such that $\sigma x(t) = x(t + T_\sigma)$. The relative period of $x(t)$ is the least $T > 0$ such that $x(T) \in \Sigma x_0$.

If $\dim \Sigma = \dim \Gamma$, then generically P is hyperbolic, hence isolated, the stability of P is determined by its

Floquet exponents, and P persists under perturbation as a periodic solution with spatial symmetry Δ and spatiotemporal symmetry Σ . For Γ infinite and $N(\Delta)/\Delta$ finite, generically P is isolated in $\text{Fix } \Delta$ and the neutral Floquet exponent has multiplicity $\dim \Gamma - \dim \Sigma + 1$.

Relative Periodic Solutions

A solution $x(t)$ is a relative periodic solution if it is not a relative equilibrium and $x(T) \in \Gamma x(0)$ for some $T > 0$. The least such T is the relative period. The spatial symmetry group $\Delta = \Sigma_{x(t)}$ for some, hence all, t . The spatiotemporal symmetry group Σ is the closed subgroup of Γ generated by Δ and σ , where $x(T) = \sigma x(0)$, and generically $\Sigma/\Delta \cong \mathbb{T}^d \times \mathbb{Z}_q$ is a maximal topologically cyclic (Cartan) subgroup of $N(\Delta)/\Delta$ containing $\sigma\Delta$. Then $x(t)$ is a $(d+1)$ -dimensional quasiperiodic motion.

The dynamics near the relative periodic solution is again governed by a skew product. There exists $n \geq 1$ such that $\sigma^n = \exp(n\xi)$, where $\xi \in LZ(\Sigma)$ and $Z(\Sigma) \subset \Gamma$ is the centralizer of Σ . Define $\alpha = \exp(-\xi)\sigma$. Form a semidirect product $\Delta \rtimes \mathbb{Z}_{2n}$ by adjoining to Δ an element Q of order $2n$ such that $Q\delta Q^{-1} = \sigma\delta\sigma^{-1}$ for $\delta \in \Delta$.

In a comoving frame with velocity ξ , a neighborhood of the relative periodic orbit is Γ -equivariantly diffeomorphic to $(\Gamma \times X \times S^1)/\Delta \rtimes \mathbb{Z}_{2n}$, where X is a $\Delta \rtimes \mathbb{Z}_{2n}$ -invariant cross section, $S^1 = \mathbb{R}/2n\mathbb{Z}$ and $\Delta \rtimes \mathbb{Z}_{2n}$ acts on $\Gamma \times X \times S^1$ as

$$\begin{aligned}\delta \cdot (\gamma, x, \theta) &= (\gamma\delta^{-1}, \delta x, \theta) \\ Q \cdot (\gamma, x, \theta) &= (\gamma\alpha^{-1}, Qx, \theta + 1)\end{aligned}$$

The Γ -equivariant ODE on $(\Gamma \times X \times S^1)/\Delta \rtimes \mathbb{Z}_{2n}$ lifts to a $\Gamma \times (\Delta \rtimes \mathbb{Z}_{2n})$ -equivariant skew product

$$\dot{\gamma} = \gamma\xi(x, \theta), \quad \dot{x} = b(x, \theta), \quad \dot{\theta} = 1 \quad [5]$$

where $\xi: X \times S^1 \rightarrow L\Gamma$, $b: X \times S^1 \rightarrow X$ satisfy appropriate $\Delta \rtimes \mathbb{Z}_{2n}$ -equivariance conditions.

Robust Heteroclinic Cycles

Heteroclinic cycles, degenerate in systems without symmetry, arise robustly in equivariant systems. Let $x_1, \dots, x_m \in \mathbb{R}^n$ be saddles with $W^u(x_i) - \{x_i\} \subset \Gamma W^s(x_{i+1})$ (where $m+1=1$). If $\Sigma_1, \dots, \Sigma_m \subset \Gamma$ are isotropy subgroups, $W^u(x_i) \subset \text{Fix } \Sigma_i$, and x_{i+1} is a sink in $\text{Fix } \Sigma_i$, then saddle-sink connections from x_i to x_{i+1} persist for nearby Γ -equivariant flows. The union $\bigcup_{i=1}^m \Gamma W^u(x_i)$ forms a robust heteroclinic cycle (see the subsection ‘‘Dynamics’’ for an example). Such cycles, when asymptotically stable, are a mechanism for intermittency or bursting, notably in rotating Rayleigh-Bénard convection (where rolls disappear

and reorient themselves at approximately 60°), and provide a possible intrinsic explanation for irregular reversals of the Earth’s magnetic field.

Asymmetric perturbations (deterministic or noisy) destroy the cycles, but the perturbed attractors inherit the bursting behavior.

Establishing the existence of heteroclinic connections is often straightforward when $\dim \text{Fix } \Sigma_i = 2$ and nontrivial with $\dim \text{Fix } \Sigma_i \geq 3$. Criteria for asymptotic stability of heteroclinic cycles are given in terms of real parts of eigenvalues of $(df)_{x_i}$, and depend on the geometry of the representation of Γ .

Robust cycles exist also between more complicated dynamical states such as periodic solutions or chaotic sets (cycling chaos). When $W^u(x_i)$ connects to two or more distinct states, the collection of unstable manifolds forms a heteroclinic network leading to competition between various subnetworks.

Symmetric Attractors

Suppose that Γ is a finite group acting linearly on \mathbb{R}^n . A closed subset $A \subset \mathbb{R}^n$ has symmetry groups $\Delta = \{\gamma \in \Gamma: \gamma x = x \text{ for all } x \in A\}$, $\Sigma = \{\gamma \in \Gamma: \gamma A = A\}$. Here, Δ is an isotropy subgroup and $\Delta \subset \Sigma \subset N(\Delta)$. In applications, Δ corresponds to instantaneous symmetry and Σ to symmetry on average.

If A is an attractor (a Lyapunov stable ω -limit set) for a Γ -equivariant vector field $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$, then Σ fixes a connected component of $\text{Fix } \Delta - L$, where L is the union of proper fixed-point spaces in $\text{Fix } \Delta$.

Provided $\dim \text{Fix } \Delta \geq 3$, all pairs Δ, Σ satisfying the above restrictions arise as symmetry groups of a nonperiodic attractor A . If $\dim \text{Fix } \Delta \geq 5$, then A is realized by a uniformly hyperbolic (Axiom A) attractor.

If $\dim \text{Fix } \Delta \geq 3$ and Σ fixes a connected component of $\text{Fix } \Delta - L$, then A is realized by a periodic sink provided Σ/Δ is cyclic. If $\dim \text{Fix } \Delta = 2$, then in addition either $\Sigma = \Delta$ or $\Sigma = N(\Delta)$.

Suppose A is an attractor and $\gamma \in \Gamma - \Sigma$. Then $\gamma A \cap A = \emptyset$. Varying a parameter, A may undergo a symmetry-increasing bifurcation: A grows until it collides with γA producing a larger attractor with symmetry on average generated by Σ and γ .

Determining symmetries of an attractor by inspection is often infeasible. A detective is a Γ -equivariant polynomial $\phi: \mathbb{R}^n \rightarrow V$ where every subgroup of Γ is an isotropy subgroup for the action on V , and each component of ϕ is nonzero. Suppose that $A \subset \mathbb{R}^n$ is an attractor with physical (Sinai-Ruelle-Bowen) measure μ . By ergodicity, the time average

$$\psi_A = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \phi(x(t)) dt \in V$$

is well defined for almost every trajectory $x(t)$ in $\text{supp } \mu$. Generically, $\Sigma_{\psi_A} = \Sigma_A$ so computing the symmetry of A reduces to computing the symmetry of a point.

If Γ is an infinite compact Lie group, and A is an ω -limit set containing points of trivial isotropy, then A cannot be uniformly hyperbolic. Hence partially hyperbolic flows arise naturally in systems with continuous symmetry. Consider the skew product [4] where $\Sigma = 1$ and $h: X \rightarrow X$ possesses a hyperbolic basic set $\Lambda \subset X$ with equilibrium measure μ (for a Hölder potential). Let ν denote Haar measure on Γ . Then $\Lambda \times \Gamma$ is partially hyperbolic, and $\mu \times \nu$ is ergodic (even Bernoulli) for an open dense set of equivariant flows. Such stably ergodic flows possess strong statistical properties (rapid decay of correlations, central-limit theorem); a possible explanation for hypermeander (Brownian-like motion) of spiral waves in planar excitable media.

Forced Symmetry Breaking

In applications, symmetry is not perfect and account should be taken of Γ' -equivariant perturbations of [1] for Γ' a subgroup of Γ (including $\Gamma' = 1$). This topic is not discussed in this article, except in the subsections “Robust heteroclinic cycles” and “Branching patterns and finite determinacy.”

Equivariant Bifurcation Theory

Consider families of ODEs $\dot{x} = f(x, \lambda)$, with bifurcation parameter $\lambda \in \mathbb{R}$ and vector field $f: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$ satisfying $f(0, 0) = 0$ and the Γ -equivariance condition

$$f(\gamma x, \lambda) = \gamma f(x, \lambda) \text{ for all } x \in \mathbb{R}^n, \lambda \in \mathbb{R}, \gamma \in \Gamma$$

A local bifurcation from the equilibrium $x = 0$ occurs if $(df)_{0,0}$ is nonhyperbolic. The center subspace E^c is the sum of generalized eigenspaces corresponding to eigenvalues on the imaginary axis, and is Γ -invariant. By center manifold theory, local dynamics $((x, \lambda) \text{ near } (0, 0))$ are captured by the center manifold W^c . After center manifold reduction (or Lyapunov–Schmidt reduction if the focus is on equilibria), it may be assumed that $\mathbb{R}^n = E^c$.

If $(df)_{0,0}$ possesses zero eigenvalues, then there is a steady-state bifurcation. Generically, $(df)_{0,0} = 0$ and E^c is absolutely irreducible. There are two subcases.

If Γ acts trivially on \mathbb{R}^n , then $n = 1$ and generically there is a saddle-node (or limit point) bifurcation where the zero sets of $f(x, \lambda)$ and $\pm x^2 \pm \lambda$ are diffeomorphic for (x, λ) near $(0, 0)$. Higher-order

degeneracies can be treated using singularity theory. The equilibria and their stability determines the local dynamics. All bifurcating equilibria have isotropy Γ , so there is no symmetry breaking.

From now on, consider the remaining subcase where Γ acts absolutely irreducibly and nontrivially on \mathbb{R}^n . Then $\text{Fix } \Gamma = \{0\}$, $f(0, \lambda) \equiv 0$, and $(df)_{0,\lambda} = c(\lambda)I_n$ where generically $c'(0) \neq 0$. Assume that $c'(0) > 0$, so the “trivial solution” $x = 0$ is asymptotically stable subcritically ($\lambda < 0$) and unstable supercritically ($\lambda > 0$). Bifurcating solutions lie outside $\text{Fix } \Gamma$ and hence there is spontaneous symmetry breaking.

Axial Isotropy Subgroups

The “equivariant branching lemma” guarantees branches of equilibria with isotropy Σ for each axial isotropy subgroup. There are three associated branching patterns, see Figure 1.

If $N(\Sigma)/\Sigma = \mathbb{Z}_2$, then f_Σ is odd. Generically, $\partial_x^3 f_\Sigma(0, 0) \neq 0$, since $(x_1^2 + \cdots + x_n^2)x$ is Γ -equivariant, and there are two branches of equilibria bifurcating supercritically or subcritically together, and lying on the same group orbit. The branches form a symmetric pitchfork whose direction of branching is determined by $\text{sgn } \partial_x^3 f_\Sigma(0, 0)$.

If $N(\Sigma)/\Sigma \cong 1$, then generically f_Σ is even. If all quadratic Γ -equivariant maps vanish on $\text{Fix } \Sigma$, then the bifurcation is sub/supercritical depending on $\text{sgn } \partial_x^3 f_\Sigma(0, 0)$ but the branches lie on distinct group orbits. This is an asymmetric pitchfork.

If $\partial_x^2 f_\Sigma(0, 0) \neq 0$, then the equilibria exist transcritically: for $\lambda < 0$ and $\lambda > 0$.

The natural actions of D_m on \mathbb{R}^2 are absolutely irreducible. The axial branches are symmetric pitchforks for $m \geq 4$ even, asymmetric pitchforks for $m \geq 5$ odd, and transcritical for $m = 3$.

The actions of $D_m, m \geq 5$ odd, provide the simplest instances of hidden symmetries, where certain $N(\Sigma)/\Sigma$ -equivariant mappings on $\text{Fix } \Sigma$ do not extend to smooth Γ -equivariant mappings on \mathbb{R}^n .

Nonaxial Maximal Isotropy Subgroups

For Σ a real maximal isotropy subgroup, $\dim \text{Fix } \Sigma$ odd, there exist branches of equilibria with isotropy

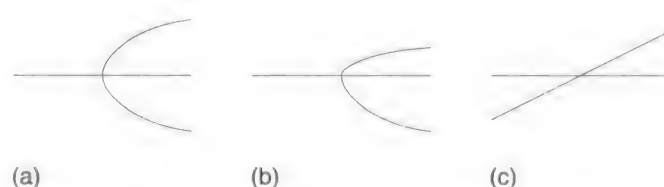


Figure 1 Axial branches: (a) supercritical symmetric pitchfork, (b) supercritical asymmetric pitchfork, and (c) transcritical branches.

Σ . When $\dim \text{Fix } \Sigma$ is even, there are examples where equilibria exist and examples where no equilibria exist. For Σ complex or quaternionic, there exist branches of rotating waves with isotropy Σ . In the quaternionic case, the rotating waves foliate the $SU(2)$ group orbits according to the Hopf fibration.

Submaximal Isotropy Subgroups

It has been conjectured falsely that steady-state bifurcation leads generically to equilibria only with maximal isotropy. The simplest counterexample is the 24-element group $\Gamma = Z_3 Z_2^3$ generated by

$$\rho = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad \kappa = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

(Alternatively, $\Gamma = T \oplus Z_2(-I_3)$, where $T \subset SO(3)$ is the tetrahedral group.)

The isotropy subgroup $\Sigma = Z_2(\kappa)$ has two-dimensional fixed-point subspace $\text{Fix } \Sigma = \{(x, y, 0)\}$. The only one-dimensional fixed-point spaces contained in $\text{Fix } \Sigma$ are the x - and y -axes. The general Γ -equivariant vector field is

$$\begin{aligned} \dot{x} &= g(x^2, y^2, z^2, \lambda)x \\ \dot{y} &= g(y^2, z^2, x^2, \lambda)y \\ \dot{z} &= g(z^2, x^2, y^2, \lambda)z \end{aligned}$$

After scaling,

$$\begin{aligned} g(x^2, y^2, z^2, \lambda) \\ = \lambda - x^2 - ay^2 - bz^2 + o(x^2, y^2, z^2, \lambda) \end{aligned} \quad [6]$$

Restricting to $\text{Fix } \Sigma$ and dividing out the axial solutions $x=0$ and $y=0$ yields at lowest order the equations $\lambda = x^2 + ay^2 = y^2 + bx^2$. Submaximal solutions exist provided $\text{sgn}(a-1) = \text{sgn}(b-1)$.

In general, the existence of equilibria with submaximal isotropy must be treated on a case-by-case basis (for each absolutely irreducible representation of Γ and isotropy subgroup Σ).

Asymptotic Stability

Subcritical and axial transcritical branches are automatically unstable. Moreover, the existence of a quadratic Γ -equivariant mapping $q: \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $x \in \text{Fix } \Sigma$ such that $(dq)_x$ has eigenvalues with nonzero real part guarantees that branches of equilibria with axial isotropy Σ are generically unstable (even when $q|_{\text{Fix } \Sigma} \equiv 0$).

There are no general results for asymptotic stability, and calculations must be done on a case-by-case basis. (The remarks in the subsection "Equilibria" are useful here.)

Branching Patterns and Finite Determinacy

The following notion of finite determinacy is based on equivariant transversality theory. Assume Γ acts absolutely irreducibly. Consider the set \mathcal{F} of Γ -equivariant vector fields $f: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$ satisfying $(df)_{0,0} = 0$. For an open dense subset of \mathcal{F} , branches of relative equilibria near $(0,0)$ are normally hyperbolic. The collection of branches of relative equilibria, together with their isotropy type, direction of branching, and stability properties, is called a branching pattern. These persist under small perturbations and are finitely determined: there exist $q = q_\Gamma \geq 2$ and an open dense subset $\mathcal{U}(q) \subset \mathcal{F}$ such that the branching patterns of f and $f+g$ are identical for $f \in \mathcal{U}(q)$, $g \in \mathcal{F}$, provided $g(x, \lambda) = o(\|x\|^q)$.

Furthermore, branching patterns are strongly finitely determined: there exist $d \geq 2$ and an open dense subset $\mathcal{S}(d) \subset \mathcal{F}$ such that the branching patterns of f and $f+g$ are identical for $f \in \mathcal{S}(d)$ and all (not necessarily equivariant) g satisfying $g(x, \lambda) = o(\|x\|^d)$.

For example, consider the hyperoctahedral group $S_n Z_2^n$, $n \geq 1$. Here S_n acts by permutations of the coordinates (x_1, \dots, x_n) and Z_2^n consists of diagonal matrices with entries ± 1 . Let $\Gamma = T Z_2^n$, where $T \subset S_n$ is a transitive subgroup. Then Γ acts absolutely irreducibly on \mathbb{R}^n and is strongly 3-determined. Submaximal branches of equilibria exist except when $T = S_n$, $T = A_n$ and, if $n=6$, $T = \text{PGL}_2(\mathbb{F}_5)$.

Dynamics

Absolutely irreducible representations have arbitrarily high dimension, so steady-state bifurcation leads to rich dynamics. The group $\Gamma = Z_3 Z_2^3$ with $\text{sgn}(a-1) \neq \text{sgn}(b-1)$ and $a+b > 2$ in [6] yields asymptotically stable heteroclinic cycles with planar connections connecting equilibria in the x -, y - and z -axes (see Figure 2). In \mathbb{R}^4 , there is the possibility of instant chaos where chaotic dynamics bifurcates directly from the equilibrium 0.

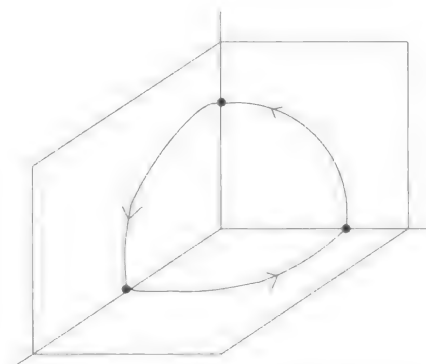


Figure 2 Robust heteroclinic cycle for the group $\Gamma = Z_3 \times Z_2^3$.

In the absence of quadratic equivariants, the invariant-sphere theorem gives an open set of equivariant vector fields for which an attracting normally hyperbolic flow-invariant $(n-1)$ -dimensional sphere bifurcates supercritically. This simplifies computations of nontrivial dynamics.

Hopf Bifurcation and Mode Interactions

Equivariant Hopf Bifurcation

The setting is the same as in the last section, except that $L = (df)_{0,0}$ has imaginary eigenvalues $\pm i\omega$ of algebraic and geometric multiplicity $n/2$. Generically, $\mathbb{R}^n = E^c$ is Γ -simple: either the direct sum of two isomorphic absolutely irreducible subspaces, or nonabsolutely irreducible.

By Birkhoff normal-form theory (see below), for any $k \geq 1$ there is a Γ -equivariant change of coordinates after which $f(x, \lambda) = f_k(x, \lambda) + o(\|x\|^k)$, where f_k is $(\Gamma \times S^1)$ -equivariant. Here $S^1 = \{\exp(tL) : t \in \mathbb{R}\}$ acts freely on \mathbb{R}^n and $\Gamma \times S^1$ acts complex irreducibly ($\mathcal{D} = \mathbb{C}$). Hence, $\dim \text{Fix } J$ is even for each isotropy subgroup $J \subset \Gamma \times S^1$, and $N(J)/J \cong S^1$ when J is maximal. The equivariant Hopf theorem guarantees, generically, branches of rotating waves with absolute period approximately $2\pi/\omega$ for each maximal isotropy subgroup J .

The notions of finite and strong finite determinacy extend to complex irreducible representations and the rotating waves persist as periodic solutions for the original Γ -equivariant vector field f . Define the spatial and spatiotemporal symmetry groups $\Delta \subset \Sigma \subset \Gamma$ as in the subsection "Periodic solutions." Then $J = \{(\sigma, \theta(\sigma)) : \sigma \in \Sigma\}$ is a twisted subgroup, with $\theta : \Sigma \rightarrow S^1$ a homomorphism and $\Delta = J \cap \Gamma = \ker \theta$.

In the non-symmetry-breaking case, where Γ acts trivially on \mathbb{R}^2 , phase-amplitude reduction leads to \mathbb{Z}_2 -equivariant amplitude equations on \mathbb{R} and higher-order degeneracies are amenable to \mathbb{Z}_2 -equivariant singularity theory. Similar comments apply to $O(2)$ -equivariant Hopf bifurcation where the amplitude equations are \mathbb{D}_4 -equivariant. The technique fails for general groups Γ .

Mode Interactions and Birkhoff Normal Form

Steady-state and Hopf bifurcations are codimension 1 and occur generically in one-parameter families of Γ -equivariant vector fields. Multiparameter families may undergo higher-codimension bifurcations called mode interactions. Suppressing parameters, steady-state/steady-state bifurcation occurs when $\mathbb{R}^n = E^c = V_1 \oplus V_2$, where V_1 and V_2 are absolutely irreducible and $L = (df)_0$ has zero eigenvalues. If V_1 and V_2 are nonisomorphic then

$L = 0$, otherwise L is nilpotent and there is an equivariant Takens–Bogdanov bifurcation. Similarly, there are codimension-2 steady-state/Hopf and Hopf/Hopf bifurcations.

Write $L = S + N$ (uniquely), where S is semisimple, N is nilpotent, and $SN = NS$. Then $\{\exp tS : t \in \mathbb{R}\}$ is a torus \mathbb{T}^p , where $p \geq 0$ is the number of rationally independent eigenvalues for L .

For each $k \geq 1$, there is a Γ -equivariant degree- k polynomial change of coordinates $P : \mathbb{R}^n \rightarrow \mathbb{R}^n$ satisfying $P(0) = 0$, $(dP)_0 = I$ transforming f to Birkhoff normal form $f_k + o(\|x\|^k)$, where f_k is $(\Gamma \times \mathbb{T}^p)$ -equivariant.

If $N \neq 0$, then $\{\exp tN^T : t \in \mathbb{R}\} \cong \mathbb{R}$ and f_k can be chosen so that the nonlinear terms are $(\Gamma \times \mathbb{T}^p \times \mathbb{R})$ -equivariant. The linear terms are not \mathbb{R} -equivariant.

The study of mode interactions proceeds by first analyzing $(\Gamma \times \mathbb{T}^p)$ -equivariant normal forms, then considering exponentially small effects of the Γ -equivariant tail. Versions of the equivariant branching lemma and equivariant Hopf theorem establish existence of certain solutions. There are numerous examples of robust heteroclinic cycles connecting (relative) equilibria and periodic solutions, symmetric chaos, and symmetry-increasing bifurcations.

Bifurcations from Relative Equilibria and Periodic Solutions

Using the skew product [4], bifurcations from a relative equilibrium with isotropy Σ for a Γ -equivariant vector field reduce to bifurcations from a fully symmetric equilibrium for a Σ -equivariant vector field h coupled with Γ drifts. If h possesses (relative) equilibria or periodic solutions, then the drift is determined generically as in the subsections "Relative equilibria and skew products" and "Relative periodic solutions." Nevertheless, solving the drift equation can be useful for understanding behavior in physical space. This is facilitated by making equivariant polynomial changes of coordinates $(\gamma Q(x), P(x))$ putting h into Birkhoff normal form and simplifying ξ .

Bifurcations from (relative) periodic solutions also reduce, mainly, to bifurcations from equilibria (with enlarged symmetry group). Based on the discussion in the subsection "Relative periodic solutions," it suffices to consider bifurcations from isolated periodic solutions $P = \{x(t)\}$ with spatial symmetry Δ and spatiotemporal symmetry Σ . Write $x(T) = \sigma x(0)$, where T is the relative period and σ is chosen so that the automorphism $\delta \mapsto \sigma^{-1} \delta \sigma$, $\delta \in \Delta$, has finite order k . Form the semidirect product $\Delta \rtimes \mathbb{Z}_{2k}$ by adjoining to Δ an element τ of order

$2k$ such that $\tau^{-1}\delta\tau = \sigma^{-1}\delta\sigma$, for $\delta \in \Delta$. Codimension-1 bifurcations from P are in one-to-one correspondence (modulo tail terms) with bifurcations from fully symmetric equilibria for a $(\Delta \rtimes \mathbb{Z}_{2k})$ -equivariant vector field. In particular, period-preserving and period-doubling bifurcations from P reduce to steady-state bifurcations, and Naimark–Sacker bifurcations reduce to Hopf bifurcations. This framework incorporates issues such as suppression of period doubling. Similar results hold for higher-codimension bifurcations.

The skew products [4] and [5] are valid for proper actions of certain noncompact Lie groups Γ provided the spatial symmetries are compact, leading to explanations of spiral and scroll wave phenomena in excitable media.

When the spatial symmetry group is noncompact, E^c may be infinite-dimensional and center manifold reduction may break down due to continuous-spectrum issues. For Euclidean symmetry, there is a theory of modulation or Ginzburg–Landau equations.

See also: Bifurcation Theory; Bifurcations in Fluid Dynamics; Bifurcations of Periodic Orbits; Central Manifolds, Normal Forms; Chaos and Attractors; Electroweak Theory; Finite Group Symmetry Breaking; Hyperbolic Dynamical Systems; Quantum Spin Systems; Quasiperiodic Systems; Singularity and Bifurcation Theory.

Further Reading

Chossat P and Lauterbach R (2000) Methods in Equivariant Bifurcations and Dynamical Systems. *Advanced Series in Nonlinear Dynamics*, vol. 15. Singapore: World Scientific.

- Crawford JD and Knobloch E (1991) Symmetry and symmetry-breaking in fluid dynamics. In: Lumley JL, Van Dyke M, and Reed HL (eds.) *Annual Review of Fluid Mechanics*, vol. 23, pp. 341–387. Palo Alto, CA: Annual Reviews.
- Fiedler B and Scheel A (2003) Spatio-temporal dynamics of reaction-diffusion patterns. In: Kirkilionis M, Krömker S, Rannacher R, and Tomi F (eds.) *Trends in Nonlinear Analysis*, pp. 23–152. Berlin: Springer.
- Field M (1996a) *Lectures on Bifurcations, Dynamics and Symmetry*. Pitman Research Notes in Mathematics Series, vol. 356. Harlow: Addison Wesley Longman.
- Field M (1996b) *Symmetry Breaking for Compact Lie Groups*. Memoirs of the American Mathematical Society, vol. 574. Providence, RI: American Mathematical Society.
- Golubitsky M and Stewart IN (2002) *The Symmetry Perspective*. Progress in Mathematics, vol. 200. Basel: Birkhäuser.
- Golubitsky M, Stewart IN, and Schaeffer D (1988) *Singularities and Groups in Bifurcation Theory, Vol. II*, Applied Mathematical Sciences, vol. 69. New York: Springer.
- Lamb JSW and Melbourne I (1999) Bifurcation from periodic solutions with spatiotemporal symmetry. In: Golubitsky M, Luss D, and Strogatz SH (eds.) *Pattern Formation in Continuous and Coupled Systems*, IMA Volumes in Mathematics and its Applications, vol. 115, pp. 175–191. New York: Springer.
- Lamb JSW, Melbourne I, and Wulff C (2003) Bifurcation from periodic solutions with spatiotemporal symmetry, including resonances and mode interactions. *Journal Differential Equations* 191: 377–407.
- Melbourne I (2000) Ginzburg–Landau theory and symmetry. In: Debnath L and Riahi DN (eds.) *Nonlinear Instability, Chaos and Turbulence, Vol 2*, Advances in Fluid Mechanics, vol. 25, pp. 79–109. Southampton: WIT Press.
- Michel L (1980) Symmetry defects and broken symmetry. Configurations. Hidden symmetry. *Reviews of Modern Physics* 52: 617–651.
- Sandstede B, Scheel A, and Wulff C (1999) Dynamical behavior of patterns with Euclidean symmetry. In: Golubitsky M, Luss D, and Strogatz SH (eds.) *Pattern Formation in Continuous and Coupled Systems*, IMA Volumes in Mathematics and its Applications, vol. 115, pp. 249–264. New York: Springer.

Symmetry and Symplectic Reduction

J-P Ortega, Université de Franche-Comté, Besançon, France

T S Ratiu, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The use of symmetries in the quantitative and qualitative study of dynamical systems has a long history that goes back to the founders of mechanics. In most cases, the symmetries of a system are used to implement a procedure generically known under the

name of “reduction” that restricts the study of its dynamics to a system of smaller dimension. This procedure is also used in a purely geometric context to construct new nontrivial manifolds having various additional structures.

Most of the reduction methods can be seen as constructions that systematize the techniques of elimination of variables found in classical mechanics. These procedures consist basically of two steps. First, one restricts the dynamics to flow-invariant submanifolds of the system in question and, second, one projects the restricted dynamics onto the symmetry orbit quotients of the spaces constructed in the first step. Sometimes, the

flow-invariant manifolds appear as the level sets of a momentum map induced by the symmetry of the system.

Symmetry Reduction

The Symmetries of a System

The standard mathematical fashion to describe the symmetries of a dynamical system (see *Dynamical Systems in Mathematical Physics: An Illustration from Water Waves*) $X \in \mathfrak{X}(M)$ defined on a manifold M denotes the Lie algebra of smooth vector fields on M endowed with the Jacobi–Lie bracket $[\cdot, \cdot]$ consists in studying its invariance properties with respect to a smooth Lie group $\Phi: G \times M \rightarrow M$ (continuous symmetries) or Lie algebra $\phi: \mathfrak{g} \rightarrow \mathfrak{X}(M)$ (infinitesimal symmetry) action. Recall that Φ is a (left) action if the map $g \in G \mapsto \Phi(g, \cdot) \in \text{Diff}(M)$ is a group homomorphism, where $\text{Diff}(M)$ denotes the group of smooth diffeomorphisms of the manifold M . The map ϕ is a (left) Lie algebra action if the map $\xi \in \mathfrak{g} \mapsto \phi(\xi) \in \mathfrak{X}(M)$ is a Lie algebra antihomomorphism and the map $(m, \xi) \in M \times \mathfrak{g} \mapsto \phi(\xi)(m) \in TM$ is smooth. The vector field X is said to be G -symmetric whenever it is equivariant with respect to the G -action Φ , that is, $X \circ \Phi_g = T\Phi_g \circ X$, for any $g \in G$. The space of G -symmetric vector fields on M is denoted by $\mathfrak{X}(M)^G$. The flow F_t of a G -symmetric vector field $X \in \mathfrak{X}(M)^G$ is G -equivariant, that is, $F_t \circ \Phi_g = \Phi_g \circ F_t$, for any $g \in G$. The vector field X is said to be \mathfrak{g} -symmetric if $[\phi(\xi), X] = 0$, for any $\xi \in \mathfrak{g}$.

If \mathfrak{g} is the Lie algebra of the Lie group G (see *Lie Groups: General Theory*) then the infinitesimal generators $\xi_M \in \mathfrak{X}(M)$ of a smooth G -group action defined by

$$\xi_M(m) := \left. \frac{d}{dt} \right|_{t=0} \Phi(\exp t\xi, m), \quad \xi \in \mathfrak{g}, \quad m \in M$$

constitute a smooth Lie algebra \mathfrak{g} -action and we denote in this case $\phi(\xi) = \xi_M$.

If $m \in M$, the closed Lie subgroup $G_m := \{g \in G \mid \Phi(g, m) = m\}$ is called the isotropy or symmetry subgroup of m . Similarly, the Lie subalgebra $\mathfrak{g}_m := \{\xi \in \mathfrak{g} \mid \phi(\xi)(m) = 0\}$ is called the isotropy or symmetry subalgebra of m . If \mathfrak{g} is the Lie algebra of G and the Lie algebra action is given by the infinitesimal generators, then \mathfrak{g}_m is the Lie algebra of G_m . The action is called free if $G_m = \{e\}$ for every $m \in M$ and locally free if $\mathfrak{g}_m = \{0\}$ for every $m \in M$. We will write interchangeably $\Phi(g, m) = \Phi_g(m) = \Phi^m(g) = g \cdot m$, for $m \in M$ and $g \in G$.

In this article we will focus mainly on continuous symmetries induced by proper Lie group actions.

The action Φ is called proper whenever for any two convergent sequences $\{m_n\}_{n \in \mathbb{N}}$ and $\{g_n \cdot m_n := \Phi(g_n, m_n)\}_{n \in \mathbb{N}}$ in M , there exists a convergent subsequence $\{g_{n_k}\}_{k \in \mathbb{N}}$ in G . Compact group actions are obviously proper.

Symmetry Reduction of Vector Fields

Let M be a smooth manifold and G a Lie group acting properly on M . Let $X \in \mathfrak{X}(M)^G$ and F_t be its (necessarily equivariant) flow. For any isotropy subgroup H of the G -action on M , the H -isotropy type submanifold $M_H := \{m \in M \mid G_m = H\}$ is preserved by the flow F_t . This property is known as the law of conservation of isotropy. The properness of the action guarantees that G_m is compact and that the (connected components of) M_H are embedded submanifolds of M for any closed subgroup H of G . The manifolds M_H are, in general, not closed in M . Moreover, the quotient group $N(H)/H$ (where $N(H)$ denotes the normalizer of H in G) acts freely and properly on M_H . Hence, if $\pi_H: M_H \rightarrow M_H/(N(H)/H)$ denotes the projection onto orbit space and $i_H: M_H \hookrightarrow M$ is the injection, the vector field X induces a unique vector field X^H on the quotient $M_H/(N(H)/H)$ defined by $X^H \circ \pi_H = T\pi_H \circ X \circ i_H$, whose flow F_t^H is given by $F_t^H \circ \pi_H = \pi_H \circ F_t \circ i_H$. We will refer to $X^H \in \mathfrak{X}(M_H/(N(H)/H))$ as the H -isotropy type reduced vector field induced by X .

This reduction technique has been widely exploited in handling specific dynamical systems. When the symmetry group G is compact and we are dealing with a linear action, the construction of the quotient $M_H/(N(H)/H)$ can be implemented in a very explicit and convenient manner by using the invariant polynomials of the action and the theorems of Hilbert and Schwarz–Mather.

Symplectic Reduction

Symplectic or Marsden–Weinstein reduction is a procedure that implements symmetry reduction for the symmetric Hamiltonian systems defined on a symplectic manifold (M, ω) . The particular case in which the symplectic manifold is a cotangent bundle is dealt with separately (see *Cotangent Bundle Reduction*). We recall that the Hamiltonian vector field $X_b \in \mathfrak{X}(M)$ associated to the Hamiltonian function $b \in C^\infty(M)$ is uniquely determined by the equality $\omega(X_b, \cdot) = db$. In this context, the symmetries $\Phi: G \times M \rightarrow M$ of interest are given by symplectic or canonical transformations, that is, $\Phi_g^* \omega = \omega$, for any $g \in G$. For canonical actions each G -invariant function $b \in C^\infty(M)^G$ has an associated G -symmetric Hamiltonian vector field X_b . A Lie

algebra action φ is called symplectic or canonical if $\mathcal{L}_{\phi(\xi)}\omega = 0$ for all $\xi \in \mathfrak{g}$, where \mathcal{L} denotes the Lie derivative operator. If the Lie algebra action is induced from a canonical Lie group action by taking its infinitesimal generators, then it is also canonical.

Momentum Maps

The symmetry reduction described in the previous section for general vector fields does not produce a well-adapted answer for symplectic manifolds (M, ω) in the sense that the reduced spaces $M_H/(N(H)/H)$ are, in general, not symplectic. To solve this problem one has to use the conservation laws associated to the canonical action, which often appear as momentum maps.

Let G be a Lie group acting canonically on the symplectic manifold (M, ω) . Suppose that for any $\xi \in \mathfrak{g}$, the vector field ξ_M is Hamiltonian, with Hamiltonian function $J^\xi \in C^\infty(M)$ and that $\xi \in \mathfrak{g} \mapsto J^\xi \in C^\infty(M)$ is linear. The map $J: M \rightarrow \mathfrak{g}^*$ defined by the relation $\langle J(z), \xi \rangle = J^\xi(z)$, for all $\xi \in \mathfrak{g}$ and $z \in M$, is called a momentum map of the G -action (see Hamiltonian Group Actions). Momentum maps, if they exist, are determined up to a constant in \mathfrak{g}^* for any connected component of M .

Examples 1

- (i) (*Linear momentum*) The phase space of an N -particle system is the cotangent space $T^*\mathbb{R}^{3N}$ endowed with its canonical symplectic structure. The additive group \mathbb{R}^3 , whose Lie algebra is abelian and is also equal to \mathbb{R}^3 , acts canonically on it by spatial translation on each factor: $v \cdot (q_i, p^i) = (q_i + v, p^i)$, with $i = 1, \dots, N$. This action has an associated momentum map $J: T^*\mathbb{R}^{3N} \rightarrow \mathbb{R}^3$, where we identified the dual of \mathbb{R}^3 with itself using the Euclidean inner product, which coincides with the classical linear momentum $J(q_i, p^i) = \sum_{i=1}^N p^i$.
- (ii) (*Angular momentum*) Let $\text{SO}(3)$ act on \mathbb{R}^3 and then, by lift, on $T^*\mathbb{R}^3$, that is, $A \cdot (q, p) = (Aq, Ap)$. This action is canonical and has as associated momentum map $J: T^*\mathbb{R}^3 \rightarrow \mathfrak{so}(3)^* \cong \mathbb{R}^3$, the classical angular momentum $J(q, p) = q \times p$.
- (iii) (*Lifted actions on cotangent bundles*) The previous two examples are particular cases of the following situation. Let $\Phi: G \times M \rightarrow M$ be a smooth Lie group action. The (left) cotangent lifted action of G on T^*Q is given by $g \cdot \alpha_q := T_{g \cdot q}^* \Phi_{g^{-1}}(\alpha_q)$ for $g \in G$ and $\alpha_q \in T^*Q$. Cotangent lifted actions preserve the canonical 1-form on T^*Q and hence are canonical. They admit an associated momentum map $J: T^*Q \rightarrow \mathfrak{g}^*$

given by $\langle J(\alpha_q), \xi \rangle = \alpha_q(\xi_Q(q))$, for any $\alpha_q \in T^*Q$ and any $\xi \in \mathfrak{g}$.

- (iv) (*Symplectic linear actions*) Let (V, ω) be a symplectic linear space and let G be a subgroup of the linear symplectic group, acting naturally on V . By the choice of G this action is canonical and has a momentum map given by $\langle J(v), \xi \rangle = (1/2)\omega(\xi_V(v), v)$, for $\xi \in \mathfrak{g}$ and $v \in V$ arbitrary.

Properties of the Momentum Map

The main feature of the momentum map that makes it of interest for use in reduction is that it encodes conservation laws for G -symmetric Hamiltonian systems. Noether's theorem states that the momentum map is a constant of the motion for the Hamiltonian vector field X_h associated to any G -invariant function $h \in C^\infty(M)^G$ (see Symmetries and Conservation Laws).

The derivative TJ of the momentum map satisfies the following two properties: $\text{range}(T_m J) = (\mathfrak{g}_m)^\circ$ and $\ker T_m J = (\mathfrak{g} \cdot m)^\omega$, for any $m \in M$, where $(\mathfrak{g}_m)^\circ$ denotes the annihilator in \mathfrak{g}^* of the isotropy subalgebra \mathfrak{g}_m of m , $\mathfrak{g} \cdot m := T_m(G \cdot m) = \{\xi_M(m) | \xi \in \mathfrak{g}\}$ is the tangent space at m to the G -orbit that contains this point, and $(\mathfrak{g} \cdot m)^\omega$ is the symplectic orthogonal space to $\mathfrak{g} \cdot m$ in the symplectic vector space $(T_m M, \omega(m))$. The first relation is sometimes called the bifurcation lemma since it establishes a link between the symmetry of a point and the rank of the momentum map at that point.

The existence of the momentum map for a given canonical action is not guaranteed. A momentum map exists if and only if the linear map $\rho: [\xi] \in \mathfrak{g}/[\mathfrak{g}, \mathfrak{g}] \mapsto [\omega(\xi_M, \cdot)] \in H^1(M, \mathbb{R})$ is identically zero. Thus, if $H^1(M, \mathbb{R}) = 0$ or $\mathfrak{g}/[\mathfrak{g}, \mathfrak{g}] = H^1(\mathfrak{g}, \mathbb{R}) = 0$ then $\rho \equiv 0$. In particular, if \mathfrak{g} is semisimple, the "first Whitehead lemma" states that $H^1(\mathfrak{g}, \mathbb{R}) = 0$ and therefore a momentum map always exists for canonical semisimple Lie algebra actions.

A natural question to ask is when the map $(\mathfrak{g}, [\cdot, \cdot]) \rightarrow (C^\infty(M), \{\cdot, \cdot\})$ defined by $\xi \mapsto J^\xi$, $\xi \in \mathfrak{g}$, is a Lie algebra homomorphism, that is, $J^{[\xi, \eta]} = [J^\xi, J^\eta]$, $\xi, \eta \in \mathfrak{g}$. Here $\{\cdot, \cdot\}: C^\infty(M) \times C^\infty(M) \rightarrow C^\infty(M)$ denotes the Poisson bracket associated to the symplectic form ω of M defined by $\{f, h\} := \omega(X_f, X_h)$, $f, h \in C^\infty(M)$. This is the case if and only if $T_z J(\xi_M(z)) = -\text{ad}_\xi^* J(z)$, for any $\xi \in \mathfrak{g}$, $z \in M$, where ad^* is the dual of the adjoint representation $\text{ad}: (\xi, \eta) \in \mathfrak{g} \times \mathfrak{g} \mapsto [\xi, \eta] \in \mathfrak{g}$ of \mathfrak{g} on itself. A momentum map that satisfies this relation is called infinitesimally equivariant. The reason behind this terminology is that this is the infinitesimal version of global or coadjoint equivariance: J is G -equivariant if $\text{Ad}_{g^{-1}}^* \circ J = J \circ \Phi_g$ or, equivalently,

$J_g^{\text{Ad}} \xi(g \cdot z) = J^\xi(z)$, for all $g \in G$, $\xi \in \mathfrak{g}$, and $z \in M$; Ad^* denotes the dual of the adjoint representation Ad of G on \mathfrak{g} . Actions admitting infinitesimally equivariant momentum maps are called Hamiltonian actions and Lie group actions with coadjoint equivariant momentum maps are called globally Hamiltonian actions. If the symmetry group G is connected then global and infinitesimal equivariance of the momentum map are equivalent concepts. If \mathfrak{g} acts canonically on (M, ω) and $H^1(\mathfrak{g}, \mathbb{R}) = \{0\}$ then this action admits at most one infinitesimally equivariant momentum map.

Since momentum maps are not uniquely defined, one may ask whether one can choose them to be equivariant. It turns out that if the momentum map is associated to the action of a compact Lie group, this can always be done. Momentum maps of cotangent lifted actions are also equivariant as are momentum maps defined by symplectic linear actions. Canonical actions of semisimple Lie algebras on symplectic manifolds admit infinitesimally equivariant momentum maps, since the “second Whitehead lemma” states that $H^2(\mathfrak{g}, \mathbb{R}) = 0$ if \mathfrak{g} is semisimple. We shall identify below a specific element of $H^2(\mathfrak{g}, \mathbb{R})$ which is the obstruction to the equivariance of a momentum map (assuming it exists).

Even though, in general, it is not possible to choose a coadjoint equivariant momentum map, it turns out that when the symplectic manifold is connected there is an affine action on the dual of the Lie algebra with respect to which the momentum map is equivariant. Define the nonequivariance 1-cocycle associated to J as the map $\sigma: G \rightarrow \mathfrak{g}^*$ given by $g \mapsto J(\Phi_g(z)) - \text{Ad}_{g^{-1}}^*(J(z))$. The connectivity of M implies that the right-hand side of this equality is independent of the point $z \in M$. In addition, σ is a (left) \mathfrak{g}^* -valued 1-cocycle on G with respect to the coadjoint representation of G on \mathfrak{g}^* , that is, $\sigma(gh) = \sigma(g) + \text{Ad}_{g^{-1}}^* \sigma(h)$ for all $g, h \in G$. Relative to the affine action $\Theta: G \times \mathfrak{g}^* \rightarrow \mathfrak{g}^*$ given by $(g, \mu) \mapsto \text{Ad}_{g^{-1}}^* \mu + \sigma(g)$, the momentum map J is equivariant. The “reduction lemma,” the main technical ingredient in the proof of the reduction theorem, states that for any $m \in M$ we have

$$\mathfrak{g}_{J(m)} \cdot m = \mathfrak{g} \cdot m \cap \ker T_m J = \mathfrak{g} \cdot m \cap (\mathfrak{g} \cdot m)^\omega$$

where $\mathfrak{g}_{J(m)}$ is the Lie algebra of the isotropy group $G_{J(m)}$ of $J(m) \in \mathfrak{g}^*$ with respect to the affine action of G on \mathfrak{g}^* induced by the nonequivariance 1-cocycle of J .

The Symplectic Reduction Theorem

The symplectic reduction procedure that we now present consists of constructing a new symplectic

manifold out of a given symmetric one in which the conservation laws encoded in the form of a momentum map and the degeneracies associated to the symmetry have been eliminated. This strategy allows the reduction of a symmetric Hamiltonian dynamical system to a dimensionally smaller one. This reduction procedure preserves the symplectic category, that is, if we start with a Hamiltonian system on a symplectic manifold, the reduced system is also a Hamiltonian system on a symplectic manifold. The reduced symplectic manifold is usually referred to as the symplectic or Marsden–Weinstein reduced space.

Theorem 2 *Let $\Phi: G \times M \rightarrow M$ be a free proper canonical action of the Lie group G on the connected symplectic manifold (M, ω) . Suppose that this action has an associated momentum map $J: M \rightarrow \mathfrak{g}^*$, with nonequivariance 1-cocycle $\sigma: G \rightarrow \mathfrak{g}^*$. Let $\mu \in \mathfrak{g}^*$ be a value of J and denote by G_μ the isotropy of μ under the affine action of G on \mathfrak{g}^* . Then:*

- (i) *The space $M_\mu := J^{-1}(\mu)/G_\mu$ is a regular quotient manifold and, moreover, it is a symplectic manifold with symplectic form ω_μ uniquely characterized by the relation*

$$\pi_\mu^* \omega_\mu = i_\mu^* \omega$$

The maps $i_\mu: J^{-1}(\mu) \hookrightarrow M$ and $\pi_\mu: J^{-1}(\mu) \rightarrow J^{-1}(\mu)/G_\mu$ denote the inclusion and the projection, respectively. The pair (M_μ, ω_μ) is called the symplectic point reduced space.

- (ii) *Let $h \in C^\infty(M)^G$ be a G -invariant Hamiltonian. The flow F_t of the Hamiltonian vector field X_h leaves the connected components of $J^{-1}(\mu)$ invariant and commutes with the G -action, so it induces a flow F_t^μ on M_μ defined by $\pi_\mu \circ F_t \circ i_\mu = F_t^\mu \circ \pi_\mu$.*
- (iii) *The vector field generated by the flow F_t^μ on (M_μ, ω_μ) is Hamiltonian with associated reduced Hamiltonian function $h_\mu \in C^\infty(M_\mu)$ defined by $h_\mu \circ \pi_\mu = h \circ i_\mu$. The vector fields X_h and X_{h_μ} are π_μ -related. The triple $(M_\mu, \omega_\mu, h_\mu)$ is called the reduced Hamiltonian system.*
- (iv) *Let $k \in C^\infty(M)^G$ be another G -invariant function. Then $\{h, k\}$ is also G -invariant and $\{h, k\}_\mu = \{h_\mu, k_\mu\}_{M_\mu}$, where $\{\cdot, \cdot\}_{M_\mu}$ denotes the Poisson bracket associated to the symplectic form ω_μ on M_μ .*

Reconstruction of Dynamics

We pose now the question converse to the reduction of a Hamiltonian system. Assume that an integral curve $c_\mu(t)$ of the reduced Hamiltonian system X_{h_μ}

on (M_μ, ω_μ) is known. Let $m_0 \in J^{-1}(\mu)$ be given. One can determine from this data the integral curve of the Hamiltonian system X_h with initial condition m_0 . In other words, one can reconstruct the solution of the given system knowing the corresponding reduced solution. The general method of reconstruction is the following. Pick a smooth curve $d(t)$ in $J^{-1}(\mu)$ such that $d(0) = m_0$ and $\pi_\mu(d(t)) = c_\mu(t)$. Then, if $c(t)$ denotes the integral curve of X_h with $c(0) = m_0$, we can write $c(t) = g(t) \cdot d(t)$ for some smooth curve $g(t)$ in G_μ that is obtained in two steps. First, one finds a smooth curve $\xi(t)$ in \mathfrak{g}_μ such that $\xi(t)_M(d(t)) = X_h(d(t)) - d(t)$. With the $\xi(t) \in \mathfrak{g}_\mu$ just obtained, one solves the nonautonomous differential equation $\dot{g}(t) = T_e L_{g(t)} \xi(t)$ on G_μ with $g(0) = e$.

The Orbit Formulation of the Symplectic Reduction Theorem

There is an alternative approach to the reduction theorem which consists of choosing as numerator of the symplectic reduced space the group invariant saturation of the level sets of the momentum map. This option produces as a result a space that is symplectomorphic to the Marsden–Weinstein quotient but presents the advantage of being more appropriate in the context of quantization problems. Additionally, this approach makes easier the comparison of the symplectic reduced spaces corresponding to different values of the momentum map which is important in the context of Poisson reduction (see Poisson Reduction). In carrying out this construction, one needs to use the natural symplectic structures that one can define on the orbits of the affine action of a group on the dual of its Lie algebra and that we now quickly review.

Let G be a Lie group, $\sigma: G \rightarrow \mathfrak{g}^*$ a coadjoint 1-cocycle, and $\mu \in \mathfrak{g}^*$. Let \mathcal{O}_μ be the orbit through μ of the affine G -action on \mathfrak{g}^* associated to σ . If $\Sigma: \mathfrak{g} \times \mathfrak{g} \rightarrow \mathbb{R}$ defined by

$$\Sigma(\xi, \eta) := \left. \frac{d}{dt} \right|_{t=0} \langle \sigma(\exp(t\xi), \eta) \rangle$$

is a real-valued Lie algebra 2-cocycle (which is always the case if σ is the derivative of a smooth real-valued group 2-cocycle or if σ is the non-equivariance 1-cocycle of a momentum map), that is, $\Sigma: \mathfrak{g} \times \mathfrak{g} \rightarrow \mathbb{R}$ is skew-symmetric and $\Sigma([\xi, \eta], \zeta) + \Sigma([\eta, \zeta], \xi) + \Sigma([\zeta, \xi], \eta) = 0$ for all $\xi, \eta, \zeta \in \mathfrak{g}$, then the affine orbit \mathcal{O}_μ is a symplectic manifold with G -invariant symplectic structure $\omega_{\mathcal{O}_\mu}^\pm$ given by

$$\omega_{\mathcal{O}_\mu}^\pm(\nu)(\xi_{\mathfrak{g}^*}(\nu), \eta_{\mathfrak{g}^*}(\nu)) = \pm \langle \nu, [\xi, \eta] \rangle \mp \Sigma(\xi, \eta) \quad [1]$$

for arbitrary $\nu \in \mathcal{O}_\mu$, and $\xi, \eta \in \mathfrak{g}$. The symbol $\xi_{\mathfrak{g}^*}(\nu) := -\text{ad}_\xi^* \nu + \Sigma(\xi, \cdot)$ denotes the infinitesimal generator of the affine action on \mathfrak{g}^* associated to $\xi \in \mathfrak{g}$. The symplectic structures $\omega_{\mathcal{O}_\mu}^\pm$ on \mathcal{O}_μ are called the (\pm) -orbit or Kostant–Kirillov–Souriau (KKS) symplectic forms.

This symplectic form can be obtained from Theorem 2 by considering the symplectic reduction of the cotangent bundle T^*G endowed with the magnetic symplectic structure $\overline{\omega}_\Sigma := \omega_{\text{can}} - \pi^* B_\Sigma$, where ω_{can} is the canonical symplectic form on T^*G , $\pi: T^*G \rightarrow G$ is the projection onto the base, and $B_\Sigma \in \Omega^2(G)^G$ is a left-invariant 2-form on G whose value at the identity is the Lie algebra 2-cocycle $\Sigma: \mathfrak{g} \times \mathfrak{g} \rightarrow \mathbb{R}$. Since Σ is a cocycle, it follows that B_Σ is closed and hence $\overline{\omega}_\Sigma$ is a symplectic form. Moreover, the lifting of the left translations on G provides a canonical G -action on T^*G that has a momentum map given by $J(g, \mu) = \Theta(g, \mu)$, $(g, \mu) \in G \times \mathfrak{g}^* \simeq T^*G$, where the trivialization $G \times \mathfrak{g}^* \simeq T^*G$ is obtained via left translations. Symplectic reduction using these ingredients yields symplectic reduced spaces that are naturally symplectically diffeomorphic to the affine orbits \mathcal{O}_μ with the symplectic form [1].

Theorem 3 (Symplectic orbit reduction). *Let $\Phi: G \times M \rightarrow M$ be a free proper canonical action of the Lie group G on the connected symplectic manifold (M, ω) . Suppose that this action has an associated momentum map $J: M \rightarrow \mathfrak{g}^*$, with nonequivariance 1-cocycle $\sigma: G \rightarrow \mathfrak{g}^*$. Let $\mathcal{O}_\mu := G \cdot \mu \subset \mathfrak{g}^*$ be the G -orbit of the point $\mu \in \mathfrak{g}^*$ with respect to the affine action of G on \mathfrak{g}^* associated to σ . Then the set $M_{\mathcal{O}_\mu} := J^{-1}(\mathcal{O}_\mu)/G$ is a regular quotient symplectic manifold with the symplectic form $\omega_{\mathcal{O}_\mu}$ uniquely characterized by the relation $i_{\mathcal{O}_\mu}^* \omega = \pi_{\mathcal{O}_\mu}^* \omega_{\mathcal{O}_\mu} + J_{\mathcal{O}_\mu}^* \omega_{\mathcal{O}_\mu}^+$, where $J_{\mathcal{O}_\mu}$ is the restriction of J to $J^{-1}(\mathcal{O}_\mu)$ and $\omega_{\mathcal{O}_\mu}^+$ is the $(+)$ -symplectic structure on the affine orbit \mathcal{O}_μ . The maps $i_{\mathcal{O}_\mu}: J^{-1}(\mathcal{O}_\mu) \hookrightarrow M$ and $\pi_{\mathcal{O}_\mu}: J^{-1}(\mathcal{O}_\mu) \rightarrow M_{\mathcal{O}_\mu}$ are natural injection and the projection, respectively. The pair $(M_{\mathcal{O}_\mu}, \omega_{\mathcal{O}_\mu})$ is called the symplectic orbit reduced space. Statements similar to (ii)–(iv) in Theorem 2 can be formulated for the orbit reduced spaces $(M_{\mathcal{O}_\mu}, \omega_{\mathcal{O}_\mu})$.*

We emphasize that given a momentum value $\mu \in \mathfrak{g}^*$, the reduced spaces M_μ and $M_{\mathcal{O}_\mu}$ are symplectically diffeomorphic via the projection to the quotients of the inclusion $J^{-1}(\mu) \hookrightarrow J^{-1}(\mathcal{O}_\mu)$.

Reduction at a general point can be replaced by reduction at zero at the expense of enlarging the manifold by the affine orbit. Consider the canonical diagonal action of G on the symplectic difference $M \ominus \mathcal{O}_\mu^+$, which is the manifold $M \times \mathcal{O}_\mu$ with the symplectic form $\pi_1^* \omega - \pi_2^* \omega_{\mathcal{O}_\mu}^+$, where $\pi_1: M \times \mathcal{O}_\mu \rightarrow M$ and $\pi_2: M \times \mathcal{O}_\mu \rightarrow \mathcal{O}_\mu$ are the projections.

A momentum map for this action is given by $J \circ \pi_1 - \pi_2 : M \ominus \mathcal{O}_\mu^+ \rightarrow \mathfrak{g}^*$. Let $(M \ominus \mathcal{O}_\mu^+)_0 := ((J \circ \pi_1 - \pi_2)^{-1}(0)/G, (\omega \ominus \omega_{\mathcal{O}_\mu^+})_0)$ be the symplectic point reduced space at zero.

Theorem 4 (Shifting theorem). *Under the hypotheses of the symplectic orbit reduction theorem (Theorem 3), the symplectic orbit reduced space $M_{\mathcal{O}_\mu}$, the point reduced spaces M_μ , and $(M \ominus \mathcal{O}_\mu^+)_0$ are symplectically diffeomorphic.*

Singular Reduction

In the previous section we carried out symplectic reduction for free and proper actions. The freeness guarantees via the bifurcation lemma that the momentum map J is a submersion and hence the level sets $J^{-1}(\mu)$ are smooth manifolds. Freeness and properness ensure that the orbit spaces $M_\mu := J^{-1}(\mu)/G_\mu$ are regular quotient manifolds. The theory of singular reduction studies the properties of the orbit space M_μ when the hypothesis on the freeness of the action is dropped. The main result in this situation shows that these quotients are symplectic Whitney stratified spaces, in the sense that the strata are symplectic manifolds in a very natural way; moreover, the local properties of this Whitney stratification make it into what is called a cone space. This statement is referred to as the “symplectic stratification theorem” and adapts to the symplectic symmetric context the stratification theorem of the orbit space of a proper Lie group action by using its orbit type manifolds. In order to present this result, we review the necessary definitions and results on stratified spaces (see Singularity and Bifurcation Theory for more information on singularity theory).

Stratified Spaces

Let \mathcal{Z} be a locally finite partition of the topological space P into smooth manifolds $S_i \subset P, i \in I$. We assume that the manifolds $S_i \subset P, i \in I$, with their manifold topology are locally closed topological subspaces of P . The pair (P, \mathcal{Z}) is a decomposition of P with pieces in \mathcal{Z} when the following condition is satisfied:

Condition (DS) If $R, S \in \mathcal{Z}$ are such that $R \cap \bar{S} \neq \emptyset$, then $R \subset \bar{S}$. In this case we write $R \preceq S$. If, in addition, $R \neq S$ we say that R is incident to S or that it is a boundary piece of S and write $R \prec S$.

The above condition is called the frontier condition and the pair (P, \mathcal{Z}) is called a decomposed space. The dimension of P is defined as $\dim P = \sup\{\dim S_i \mid S_i \in \mathcal{Z}\}$. If $k \in \mathbb{N}$, the k -skeleton P^k of P is the union of all the pieces of dimension smaller than or equal to k ; its

topology is the relative topology induced by P . The depth $\text{dp}(z)$ of any $z \in (P, \mathcal{Z})$ is defined as

$$\text{dp}(z) := \sup\{k \in \mathbb{N} \mid \exists S_0, S_1, \dots, S_k \in \mathcal{Z} \text{ with } z \in S_0 \prec S_1 \prec \dots \prec S_k\}$$

Since for any two elements $x, y \in S$ in the same piece $S \in P$ we have $\text{dp}(x) = \text{dp}(y)$, the depth $\text{dp}(S)$ of the piece S is well defined by $\text{dp}(S) := \text{dp}(x), x \in S$. Finally, the depth $\text{dp}(P)$ of (P, \mathcal{Z}) is defined by $\text{dp}(P) := \sup\{\text{dp}(S) \mid S \in \mathcal{Z}\}$.

A continuous mapping $f: P \rightarrow Q$ between the decomposed spaces (P, \mathcal{Z}) and (Q, \mathcal{Y}) is a morphism of decomposed spaces if, for every piece $S \in \mathcal{Z}$, there is a piece $T \in \mathcal{Y}$ such that $f(S) \subset T$ and the restriction $f|_S: S \rightarrow T$ is smooth. If (P, \mathcal{Z}) and (P, \mathcal{T}) are two decompositions of the same topological space we say that \mathcal{Z} is coarser than \mathcal{T} or that \mathcal{T} is finer than \mathcal{Z} if the identity mapping $(P, \mathcal{T}) \rightarrow (P, \mathcal{Z})$ is a morphism of decomposed spaces. A topological subspace $Q \subset P$ is a decomposed subspace of (P, \mathcal{Z}) if, for all pieces $S \in \mathcal{Z}$, the intersection $S \cap Q$ is a submanifold of S and the corresponding partition $\mathcal{Z} \cap Q$ forms a decomposition of Q .

Let P be a topological space and $z \in P$. Two subsets A and B of P are said to be equivalent at z if there is an open neighborhood U of z such that $A \cap U = B \cap U$. This relation constitutes an equivalence relation on the power set of P . The class of all sets equivalent to a given subset A at z will be denoted by $[A]_z$ and called the set germ of A at z . If $A \subset B \subset P$, we say that $[A]_z$ is a subgerm of $[B]_z$, and denote $[A]_z \subset [B]_z$.

A stratification of the topological space P is a map \mathcal{S} that associates to any $z \in P$ the set germ $\mathcal{S}(z)$ of a closed subset of P such that the following condition is satisfied:

Condition (ST) For every $z \in P$ there is a neighborhood U of z and a decomposition \mathcal{Z} of U such that for all $y \in U$ the germ $\mathcal{S}(y)$ coincides with the set germ of the piece of \mathcal{Z} that contains y .

The pair (P, \mathcal{S}) is called a stratified space. Any decomposition of P defines a stratification of P by associating to each of its points the set germ of the piece in which it is contained. The converse is, by definition, locally true.

The Strata

Two decompositions \mathcal{Z}_1 and \mathcal{Z}_2 of P are said to be equivalent if they induce the same stratification of P . If \mathcal{Z}_1 and \mathcal{Z}_2 are equivalent decompositions of P then, for all $z \in P$, we have that $\text{dp}_{\mathcal{Z}_1}(z) = \text{dp}_{\mathcal{Z}_2}(z)$. Any stratified space (P, \mathcal{S}) has a unique decomposition $\mathcal{Z}_{\mathcal{S}}$ associated with the following maximality property: for any open subset $U \subset P$ and any

decomposition \mathcal{Z} of P inducing \mathcal{S} over U , the restriction of \mathcal{Z}_S to U is coarser than the restriction of \mathcal{Z} to U . The decomposition \mathcal{Z}_S is called the canonical decomposition associated to the stratification (P, \mathcal{S}) . It is often denoted by \mathcal{S} and its pieces are called the strata of P . The local finiteness of the decomposition \mathcal{Z}_S implies that for any stratum S of (P, \mathcal{S}) there are only finitely many strata R with $S \prec R$. Henceforth, the symbol \mathcal{S} in the stratification (P, \mathcal{S}) will denote both the map that associates to each point a set germ and the set of pieces associated to the canonical decomposition induced by the stratification of P .

Stratified Spaces with Smooth Structure

Let (P, \mathcal{S}) be a stratified space. A singular or stratified chart of P is a homeomorphism $\phi: U \rightarrow \phi(U) \subset \mathbb{R}^n$ from an open set $U \subset P$ to a subset of \mathbb{R}^n such that for every stratum $S \in \mathcal{S}$ the image $\phi(U \cap S)$ is a submanifold of \mathbb{R}^n and the restriction $\phi|_{U \cap S}: U \cap S \rightarrow \phi(U \cap S)$ is a diffeomorphism. Two singular charts $\phi: U \rightarrow \phi(U) \subset \mathbb{R}^n$ and $\varphi: V \rightarrow \varphi(V) \subset \mathbb{R}^m$ are compatible if for any $z \in U \cap V$ there exist an open neighborhood $W \subset U \cap V$ of z , a natural number $N \geq \max\{n, m\}$, open neighborhoods $O, O' \subset \mathbb{R}^N$ of $\phi(U) \times \{0\}$ and $\varphi(V) \times \{0\}$, respectively, and a diffeomorphism $\psi: O \rightarrow O'$ such that $i_m \circ \varphi|_W = \psi \circ i_n \circ \phi|_W$, where i_n and i_m denote the natural embeddings of \mathbb{R}^n and \mathbb{R}^m into \mathbb{R}^N by using the first n and m coordinates, respectively. The notion of singular or stratified atlas is the natural generalization for stratifications of the concept of atlas existing for smooth manifolds. Analogously, we can talk of compatible and maximal stratified atlases. If the stratified space (P, \mathcal{S}) has a well-defined maximal atlas, then we say that this atlas determines a smooth or differentiable structure on P . We will refer to (P, \mathcal{S}) as a smooth stratified space.

The Whitney Conditions

Let M be a manifold and $R, S \subset M$ two submanifolds. We say that the pair (R, S) satisfies the Whitney condition (A) at the point $z \in R$ if the following condition is satisfied:

Condition (A) For any sequence of points $\{z_n\}_{n \in \mathbb{N}}$ in S converging to $z \in R$ for which the sequence of tangent spaces $\{T_{z_n} S\}_{n \in \mathbb{N}}$ converges in the Grassmann bundle of $\dim S$ -dimensional subspaces of TM to $\tau \subset T_z M$, we have that $T_z R \subset \tau$.

Let $\phi: U \rightarrow \mathbb{R}^n$ be a smooth chart of M around the point z . The Whitney condition (B) at the point

$z \in R$ with respect to the chart (U, ϕ) is given by the following statement:

Condition (B) Let $\{x_n\}_{n \in \mathbb{N}} \subset R \cap U$ and $\{y_n\}_{n \in \mathbb{N}} \subset S \cap U$ be two sequences with the same limit

$$z = \lim_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} y_n$$

and such that $x_n \neq y_n$, for all $n \in \mathbb{N}$. Suppose that the set of connecting lines $\overline{\phi(x_n)\phi(y_n)} \subset \mathbb{R}^n$ converges in projective space to a line L and that the sequence of tangent spaces $\{T_{y_n} S\}_{n \in \mathbb{N}}$ converges in the Grassmann bundle of $(\dim S)$ -dimensional subspaces of TM to $\tau \subset T_z M$. Then, $(T_z \phi)^{-1}(L) \subset \tau$.

If the condition (A) (respectively (B)) is verified for every point $z \in R$, the pair (R, S) is said to satisfy the Whitney condition (A) (respectively (B)). It can be verified that Whitney's condition (B) does not depend on the chart used to formulate it. A stratified space with smooth structure such that, for every pair of strata, Whitney's condition (B) is satisfied is called a Whitney space.

Cone Spaces and Local Triviality

Let P be a topological space. Consider the equivalence relation \sim in the product $P \times [0, \infty)$ given by $(z, a) \sim (z', a')$ if and only if $a = a' = 0$. We define the cone CP on P as the quotient topological space $P \times [0, \infty) / \sim$. If P is a smooth manifold then the cone CP is a decomposed space with two pieces, namely, $P \times (0, \infty)$ and the vertex which is the class corresponding to any element of the form $(z, 0)$, $z \in P$, that is, $P \times \{0\}$. Analogously, if (P, \mathcal{Z}) is a decomposed (stratified) space then the associated cone CP is also a decomposed (stratified) space whose pieces (strata) are the vertex and the sets of the form $S \times (0, \infty)$, with $S \in \mathcal{Z}$. This implies, in particular, that $\dim CP = \dim P + 1$ and $\text{dp}(CP) = \text{dp}(P) + 1$.

A stratified space (P, \mathcal{S}) is said to be locally trivial if for any $z \in P$ there exist a neighborhood U of z , a stratified space (F, \mathcal{S}^F) , a distinguished point $0 \in F$, and an isomorphism of stratified spaces

$$\psi: U \rightarrow (S \cap U) \times F$$

where S is the stratum that contains z and ψ satisfies $\psi^{-1}(y, 0) = y$, for all $y \in S \cap U$. When F is given by a cone CL over a compact stratified space L then L is called the link of z .

An important corollary of "Thom's first isotopy lemma" guarantees that every Whitney stratified space is locally trivial. A converse to this implication needs the introduction of cone spaces. Their definition is given by recursion on the depth of the space.

Definition 5 Let $m \in \mathbb{N} \cup \{\infty, \omega\}$. A cone space of class C^m and depth 0 is the union of countably many C^m manifolds together with the stratification whose strata are the unions of the connected components of equal dimension. A cone space of class C^m and depth $d + 1, d \in \mathbb{N}$, is a stratified space (P, \mathcal{S}) with a C^m differentiable structure such that for any $z \in P$ there exists a connected neighborhood U of z , a compact cone space L of class C^m and depth d called the link, and a stratified isomorphism

$$\psi : U \rightarrow (S \cap U) \times CL$$

where S is the stratum that contains the point z , the map ψ satisfies $\psi^{-1}(y, 0) = y$, for all $y \in S \cap U$, and 0 is the vertex of the cone CL .

If $m \neq 0$ then L is required to be embedded into a sphere via a fixed smooth global singular chart $\varphi : L \rightarrow S^l$ that determines the smooth structure of CL . More specifically, the smooth structure of CL is generated by the global chart $\tau : [z, t] \in CL \mapsto t\varphi(z) \in \mathbb{R}^{l+1}$. The maps $\psi : U \rightarrow (S \cap U) \times CL$ and $\varphi : L \rightarrow S^l$ are referred to as a cone chart and a link chart, respectively. Moreover, if $m \neq 0$ then ψ and ψ^{-1} are required to be differentiable of class C^m as maps between stratified spaces with a smooth structure.

The Symplectic Stratification Theorem

Let (M, ω) be a connected symplectic manifold acted canonically and properly upon by a Lie group G . Suppose that this action has an associated momentum map $J : M \rightarrow \mathfrak{g}^*$ with nonequivariance 1-cocycle $\sigma : G \rightarrow \mathfrak{g}^*$. Let $\mu \in \mathfrak{g}^*$ be a value of J, G_μ the isotropy subgroup of μ with respect to the affine action $\Theta : G \times \mathfrak{g}^* \rightarrow \mathfrak{g}^*$ determined by σ , and let $H \subset G$ be an isotropy subgroup of the G -action on M . Let M_H^z be the connected component of the H -isotropy type manifold that contains a given element $z \in M$ such that $J(z) = \mu$ and let $G_\mu M_H^z$ be its G_μ -saturation. Then the following hold:

1. The set $J^{-1}(\mu) \cap G_\mu M_H^z$ is a submanifold of M .
2. The set $M_\mu^{(H)} := [J^{-1}(\mu) \cap G_\mu M_H^z] / G_\mu$ has a unique quotient differentiable structure such that the canonical projection $\pi_\mu^{(H)} : J^{-1}(\mu) \cap G_\mu M_H^z \longrightarrow M_\mu^{(H)}$ is a surjective submersion.
3. There is a unique symplectic structure $\omega_\mu^{(H)}$ on $M_\mu^{(H)}$ characterized by

$$i_\mu^{(H)*} \omega = \pi_\mu^{(H)*} \omega_\mu^{(H)}$$

where $i_\mu^{(H)} : J^{-1}(\mu) \cap G_\mu M_H^z \hookrightarrow M$ is the natural inclusion. The pairs $(M_\mu^{(H)}, \omega_\mu^{(H)})$ will be called singular symplectic point strata.

4. Let $h \in C^\infty(M)^G$ be a G -invariant Hamiltonian. Then the flow F_t of X_h leaves the connected components of $J^{-1}(\mu) \cap G_\mu M_H^z$ invariant and commutes with the G_μ -action, so it induces a flow F_t^μ on $M_\mu^{(H)}$ that is characterized by $\pi_\mu^{(H)} \circ F_t \circ i_\mu^{(H)} = F_t^\mu \circ \pi_\mu^{(H)}$.
5. The flow F_t^μ is Hamiltonian on $M_\mu^{(H)}$, with reduced Hamiltonian function $h_\mu^{(H)} : M_\mu^{(H)} \rightarrow \mathbb{R}$ defined by $h_\mu^{(H)} \circ \pi_\mu^{(H)} = h \circ i_\mu^{(H)}$. The vector fields X_h and $X_{h_\mu^{(H)}}$ are $\pi_\mu^{(H)}$ -related.
6. Let $k : M \rightarrow \mathbb{R}$ be another G -invariant function. Then $\{h, k\}$ is also G -invariant and $\{h, k\}_\mu^{(H)} = \{h_\mu^{(H)}, k_\mu^{(H)}\}_{M_\mu^{(H)}}$, where $\{, \}_{M_\mu^{(H)}}$ denotes the Poisson bracket induced by the symplectic structure on $M_\mu^{(H)}$.

Theorem 6 (Symplectic stratification theorem). *The quotient $M_\mu := J^{-1}(\mu) / G_\mu$ is a cone space when considered as a stratified space with strata $M_\mu^{(H)}$.*

As was the case for regular reduction, this theorem can also be formulated from the orbit reduction point of view. Using that approach one can conclude that the orbit reduced spaces $M_{\mathcal{O}_\mu}$ are cone spaces symplectically stratified by the manifolds $M_{\mathcal{O}_\mu}^{(H)} := G \cdot (J^{-1}(\mu) \cap M_H^z) / G$ that have symplectic structure uniquely determined by the expression

$$i_{\mathcal{O}_\mu}^{(H)*} \omega = \pi_{\mathcal{O}_\mu}^{(H)*} \omega_{\mathcal{O}_\mu}^{(H)} + J_{\mathcal{O}_\mu}^{(H)*} \omega_{\mathcal{O}_\mu}^+$$

where $i_{\mathcal{O}_\mu}^{(H)} : G \cdot (J^{-1}(\mu) \cap M_H^z) \hookrightarrow M$ is the inclusion, $J_{\mathcal{O}_\mu}^{(H)} : G \cdot (J^{-1}(\mu) \cap M_H^z) \rightarrow \mathcal{O}_\mu$ is obtained by restriction of the momentum map J , and $\omega_{\mathcal{O}_\mu}^+$ is the $(+)$ -symplectic form on \mathcal{O}_μ . Analogous statements to (7)–(6) above with obvious modifications are valid.

See also: Cotangent Bundle Reduction; Dynamical Systems in Mathematical Physics: An Illustration from Water Waves; Graded Poisson Algebras; Hamiltonian Group Actions; Lie Groups: General Theory; Poisson Reduction; Singularity and Bifurcation Theory; Symmetries and Conservation Laws.

Further Reading

Abraham R and Marsden JE (1978) *Foundations of Mechanics*, 2nd edn. Reading, MA: Addison-Wesley.

Arms JM, Cushman R, and Gotay MJ (1991) A universal reduction procedure for Hamiltonian group actions. In: Ratiu TS (ed.) *The Geometry of Hamiltonian Systems*, pp. 33–51. New York: Springer.

Cendra H, Marsden JE, and Ratiu TS (2001) Lagrangian reduction by stages. *Memoirs of the American Mathematical Society*, Volume 152, No. 722.

Huebschmann J (2001) Singularities and Poisson geometry of certain representation spaces. In: Landsman NP, Pflaum M, and Schlichenmaier M (eds.) *Quantization of Singular Symplectic Quotients*, Progr. Math., vol. 198, pp. 119–135. Boston, MA: Birkhäuser.

- Kazhdan D, Kostant B, and Sternberg S (1978) Hamiltonian group actions and dynamical systems of Calogero type. *Communications in Pure and Applied Mathematics* 31: 481–508.
- Kirillov AA (1976) *Elements of the Theory of Representations*, Grundlehren Math. Wiss. Berlin–New York: Springer.
- Kostant B (1970) *Quantization and Unitary Representations*, Lecture Notes in Math., vol. 570, pp. 177–306. Berlin: Springer.
- Lerman E, Montgomery R, and Sjamaar R (1993) Examples of singular reduction. In: *Symplectic Geometry*, London Math. Soc. Lecture Note Ser., vol. 192, pp. 127–155. Cambridge: Cambridge University Press.
- Marsden J (1981) *Lectures on Geometric Methods in Mathematical Physics*. Philadelphia: SIAM.
- Marsden JE, Misiolek G, Ortega J-P, Perlmutter M, and Ratiu TS (2005) *Hamiltonian Reduction by Stages*, Lecture Notes in Mathematics. Berlin: Springer.
- Marsden JE and Ratiu TS (2003) *Introduction to Mechanics and Symmetry*, 2nd edn., second printing (1st edn (1994)), Texts in Applied Mathematics, vol. 17. New York: Springer.
- Marsden JE and Weinstein A (1974) Reduction of symplectic manifolds with symmetry. *Reports on Mathematical Physics* 5(1): 121–130.
- Marsden JE and Weinstein A (2001) Comments on the history, theory, and applications of symplectic reduction. In: Landsman N, Pflaum M, and Schlichenmaier M (eds.) *Quantization of Singular Symplectic Quotients*, Progress in Mathematics, vol. 198, pp. 1–20. Boston: Birkhäuser.
- Mayer KR (1973) Symmetries and integrals in mechanics. In: Peixoto M (ed.) *Dynamical Systems*, pp. 259–273. New York: Academic Press.
- Ortega J-P and Ratiu TS (2003) *Momentum Maps and Hamiltonian Reduction*, Progress in Math., vol. 222. Boston: Birkhäuser.
- Pflaum MJ (2001) *Analytic and Geometric Study of Stratified Spaces*, Lecture Notes in Mathematics, vol. 1768. Berlin: Springer-Verlag.
- Sjamaar R and Lerman E (1991) Stratified symplectic spaces and reduction. *Annals of Mathematics* 134: 375–422.
- Souriau J-M (1970) *Structure des Systèmes Dynamiques*. Paris: Dunod. (English translation by Cushman RH and Tuynman GM (1997)) *Structure of Dynamical Systems*, Progress in Math., vol. 149. Boston: Birkhäuser.

Symmetry Breaking in Field Theory

T W B Kibble, Imperial College, London, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Spontaneous symmetry breaking in its simplest form occurs when there is a symmetry of a dynamical system that is not manifest in its ground state or equilibrium state. It is a common feature of many classical and quantum systems. In quantum field theories, in the infinite-volume limit, there are new features, the appearance of unitarily inequivalent representations of the canonical commutation relations, and the possibility of a true phase transition – a point in the phase space where the thermodynamic free energy is nonanalytic. The spontaneous breaking of a continuous global symmetry implies the existence of massless particles, the Goldstone bosons, while in the local-symmetry case some or all of these may be eliminated by the Higgs mechanism. Spontaneous symmetry breaking in gauge theories is however a more elusive concept.

Breaking of Global Symmetries

In a quantum-mechanical system a (time-independent) symmetry is represented by a unitary operator \hat{U} acting on the Hilbert space of quantum states which

commutes with the Hamiltonian \hat{H} . If the ground state $|0\rangle$ of the system is not invariant under \hat{U} , then $|0'\rangle = \hat{U}|0\rangle \neq c|0\rangle$ is also a ground state. In other words, the ground state is degenerate.

For a system with a finite number of degrees of freedom, whose states are represented by vectors in a separable Hilbert space \mathcal{H} , symmetry breaking of an abelian symmetry group G is impossible, unless there are additional accidental symmetries. Consider, for example, a particle in a double-well potential

$$V = \frac{m\omega^2}{4a^2}(x^2 - a^2)^2 \quad [1]$$

which has the discrete symmetry group $G = \mathbb{Z}_2$; the inversion symmetry operator \hat{U} satisfies $\hat{U}^2 = \hat{1}$. There are then two approximate ground states $|0\rangle$ and $|0'\rangle = \hat{U}|0\rangle$, with wave functions proportional to $\exp[-(1/2)m\omega(x \mp a)^2]$. However, there is an overlap between these, and the off-diagonal matrix element $\langle 0|\hat{H}|0'\rangle$ is nonzero, although exponentially small, so the true energy eigenstates are, approximately, $|0_{\pm}\rangle = (1/\sqrt{2})(|0\rangle \pm |0'\rangle)$. (More accurate energy eigenfunctions and eigenvalues may be found by using the WKB approximation.)

Of course, if the symmetry group is nonabelian, and the ground state belongs to a nontrivial representation, then degeneracy is unavoidable. For example, if G is the rotation group $\text{SO}(3)$ (or $\text{SU}(2)$)

and the ground state has angular momentum $j \neq 0$, then it is $(2j+1)$ -fold degenerate.

The situation is different, however, in a quantum field theory. In the infinite-volume limit, even abelian symmetries can be spontaneously broken. Take, for example, a real scalar field with Lagrangian

$$\mathcal{L} = \frac{1}{2} \partial_\mu \phi \partial^\mu \phi - V = \frac{1}{2} \dot{\phi}^2 - \frac{1}{2} (\nabla \phi)^2 - V \quad [2]$$

(where we set $c = \hbar = 1$), again with a double-well potential

$$V = \frac{1}{8} \lambda (\phi^2 - \eta^2)^2 \quad [3]$$

exhibiting a \mathbb{Z}_2 symmetry under which $\phi(x) \mapsto -\phi(x)$.

At least in the semiclassical or tree approximation, there are two degenerate vacuum states $|0\rangle$ and $|0'\rangle$, with

$$\langle 0 | \dot{\phi}(x) | 0 \rangle \approx \eta \quad \text{and} \quad \langle 0' | \dot{\phi}(x) | 0' \rangle \approx -\eta \quad [4]$$

If we quantize the system in a box of finite volume \mathcal{V} , then, as earlier, there is an off-diagonal matrix element of the Hamiltonian connecting the two states, so the true ground state is (approximately) $(1/\sqrt{2})(|0\rangle + |0'\rangle)$. However, this matrix element goes to zero exponentially as $\mathcal{V} \rightarrow 0$. Even for large but finite volume, the rate of transitions from $|0\rangle$ to $|0'\rangle$ is exponentially slow.

Similarly, we can consider a complex scalar field theory with a sombrero potential:

$$\begin{aligned} \mathcal{L} &= |\dot{\phi}|^2 - |\nabla \phi|^2 - V \\ V &= \frac{1}{2} \lambda \left(|\phi|^2 - \frac{1}{2} \eta^2 \right)^2 \end{aligned} \quad [5]$$

This model is invariant under the $U(1)$ group of phase transformations, $\phi(x) \mapsto \phi(x)e^{i\alpha}$, so we now have a continuously infinite set of degenerate vacuum states $|0_\alpha\rangle$ labeled by an angle α , and satisfying

$$\langle 0_\alpha | \hat{\phi}(x) | 0_\alpha \rangle \approx \frac{1}{\sqrt{2}} \eta e^{i\alpha} \quad [6]$$

Once again, one finds that in the infinite-volume limit there are no matrix elements connecting the different vacuum states. Moreover, in this limit no polynomial formed from the field operators $\hat{\phi}(x)$ in a finite volume can have nonzero matrix elements between $|0_\alpha\rangle$ and $|0_\beta\rangle$ for $\alpha \neq \beta$. Applying the operators $\hat{\phi}(x)$ to any one of these vacuum states $|0_\alpha\rangle$, we can construct a Fock space \mathcal{H}_α , and the representations of the canonical commutation relations on these separate Hilbert spaces are unitarily inequivalent. Formally, we can introduce operators \hat{U}_α that perform the symmetry transformations:

$$\hat{U}_\alpha \hat{\phi}(x) \hat{U}_\alpha^{-1} = \hat{\phi}(x) e^{i\alpha} \quad [7]$$

However, these are not unitary operators on the spaces \mathcal{H}_β , but rather maps from one space to another: $\hat{U}_\alpha: \mathcal{H}_\beta \rightarrow \mathcal{H}_{\alpha+\beta}$ – or, alternatively, operators on the nonseparable Hilbert space $\mathcal{H} = \bigoplus_\alpha \mathcal{H}_\alpha$.

So far, our discussion has been restricted to the tree approximation. For a full quantum treatment, $V(\phi)$ must be replaced by the effective potential $V_{\text{eff}}(\phi)$, which may be defined as the minimum value of the mean energy density in all states in which the field $\hat{\phi}$ has the uniform expectation value $\langle \hat{\phi}(x) \rangle = \phi$. V_{eff} may be computed by summing vacuum loop diagrams.

A point to note is that although the degenerate vacua $|0_\alpha\rangle$ are mathematically distinct, in the absence of any external definition of phase, they are physically identical. There is no internal observational test that will distinguish them.

Symmetry-Breaking Phase Transitions

Spontaneous symmetry breaking often occurs in the context of a phase transition. At high temperature, $T \gg \eta$, there are large fluctuations in ϕ and the central hump of the potential is unimportant. Then the equilibrium state is symmetric, with $\langle \hat{\phi} \rangle = 0$. However, as the temperature falls, it becomes less probable that the field will fluctuate over the top of the hump. It will tend to fall into the trough, and acquire a nonzero average value $\langle \hat{\phi} \rangle$ – the order parameter for the phase transition – thus breaking the symmetry. The direction of symmetry breaking (e.g., the phase of ϕ in the $U(1)$ model) is random, determined in practice by small preexisting fluctuations or interactions with the environment.

One way of studying this process is to compute the temperature-dependent effective potential $V_{\text{eff}}(\phi, T)$. In the one-loop approximation, at high temperature, the leading corrections to the zero-temperature effective potential $V_{\text{eff}}(\phi, T)$ are of the form

$$\begin{aligned} V_{\text{eff}}(\phi, T) &= V_{\text{eff}}(\phi, 0) - \frac{\pi^2}{90} N_* T^4 \\ &\quad + \frac{1}{24} M_*^2(\phi) T^2 + \mathcal{O}(T) \end{aligned} \quad [8]$$

where N_* is the total number of helicity states of light particles (those with masses $\ll T$), and M_*^2 , which depends on ϕ , is the sum of their squared masses. (Fermions if present contribute to N_* with a factor of $7/8$ and to M_*^2 with a factor of $1/2$.) In the simplest case, where we have only a multiplet $\phi = (\phi_a)_{a=1, \dots, N}$ of real scalar fields, $N_* = N$ and $M_*^2 = M_{aa}^2$

(summation over a implied), where the mass-squared matrix is

$$M_{ab}^2 = \frac{\partial^2 V}{\partial \phi_a \partial \phi_b} \quad [9]$$

For example, in an $O(N)$ theory, with $V = (1/8)\lambda(\phi^2 - \eta^2)^2$, where $\phi^2 = \phi_a \phi_a$, one has

$$M_{ab}^2 = \frac{1}{2}\lambda(\phi^2 - \eta^2)\delta_{ab} + \lambda\phi_a \phi_b \quad [10]$$

whence

$$V_{\text{eff}}(\phi, T) \approx \frac{1}{8}\lambda(\phi^2 - \eta^2)^2 - \frac{\pi^2}{90}NT^4 + \frac{1}{48}\lambda T^2[(N+2)\phi^2 - N\eta^2] \quad [11]$$

It is then easy to see that the minimum occurs at $\phi=0$ for $T > T_c$, where in this approximation $T_c^2 = 12\eta^2/(N+2)$, while below the critical temperature the minimum is at

$$\phi^2 = \phi_{\text{eq}}^2(T) \approx \eta^2 - \frac{N+2}{12}T^2 \quad [12]$$

As $T \rightarrow 0$, the equilibrium state approaches one of the vacuum states $|0_n\rangle$, labeled by an N -dimensional unit vector \mathbf{n} , such that $\langle 0_n | \hat{\phi} | 0_n \rangle = \eta \mathbf{n}$.

It is often convenient to introduce a classical symmetry-breaking potential. For example, in the $O(N)$ model, we may take $V_{\text{sb}} = -\mathbf{j} \cdot \hat{\phi}(x)$, where \mathbf{j} is a constant N -vector. This has the effect of tilting the potential, thus removing the degeneracy. A characteristic of spontaneous symmetry breaking is that the limits $\mathbf{j} \rightarrow 0$ and $V \rightarrow \infty$ do not commute. If (for $T < T_c$) we take the infinite-volume limit first, and then let $\mathbf{j} \rightarrow 0$, we get different equilibrium states, depending on the direction from which \mathbf{j} approaches zero; if we fix \mathbf{n} and let $\mathbf{j} = j\mathbf{n}$, $j \rightarrow 0$, then we find

$$\lim_{j \rightarrow 0} \lim_{V \rightarrow \infty} \langle \hat{\phi}(x) \rangle_{jm} = \phi_{\text{eq}}(T)\mathbf{n} \quad [13]$$

We may also regard \mathbf{j} as representing an interaction with the external environment (e.g., other fields). If such a term is present during the cooling of the system through the phase transition, it will constrain the direction of the spontaneous symmetry breaking. Note that one always arrives in this way at one of the degenerate vacua $|0_n\rangle$, not a linear combination of them.

Goldstone Bosons

The Goldstone theorem states that spontaneous breaking of any continuous global symmetry leads inevitably (except, as we discuss later, in the

presence of long-range forces) to the appearance of massless modes – the Goldstone bosons.

The proof is straightforward. Associated with any continuous symmetry there is a Noether current satisfying the continuity equation $\partial_\mu \hat{j}^\mu = 0$ and such that infinitesimal symmetry transformations are generated by the spatial integral of \hat{j}^0 . The fact that the symmetry is broken means that there is some scalar field $\hat{\phi}(x)$ whose vacuum expectation value $\langle 0 | \hat{\phi}(0) | 0 \rangle$ is not invariant under the symmetry transformation. Hence,

$$\lim_{V \rightarrow 0} i \int_V d^3x \langle 0 | [\hat{j}^0(x), \hat{\phi}(0)] | 0 \rangle|_{x^0=0} \neq 0 \quad [14]$$

Moreover, the time derivative of this integral is

$$\begin{aligned} \lim_{V \rightarrow 0} i \int_V d^3x \langle 0 | [\partial_0 \hat{j}^0(x), \hat{\phi}(0)] | 0 \rangle|_{x^0=0} \\ = -\lim_{V \rightarrow 0} i \int_{\partial V} dS_k \langle 0 | [\hat{j}^k(x), \hat{\phi}(0)] | 0 \rangle|_{x^0=0} = 0 \end{aligned} \quad [15]$$

where ∂V is the bounding surface of V . This vanishes because the surface integral is zero – in a relativistic theory, because the commutator vanishes at space-like separation, and more generally in the absence of long-range interactions because it tends rapidly to zero at large spatial separation.

Now, inserting a complete set of momentum eigenstates $|n, \mathbf{p}\rangle$ in [14], we can see that there must exist states such that $\langle n, \mathbf{p} | \hat{\phi}(0) | 0 \rangle \neq 0$, with $p^0 \rightarrow 0$ in the limit $|\mathbf{p}| \rightarrow 0$, that is, massless modes.

One can see this more directly in the $U(1)$ model above. Consider a vacuum state $|0\rangle$ such that $\langle 0 | \hat{\phi} | 0 \rangle = \eta/\sqrt{2}$ is real. Then it is useful to shift the origin of ϕ by writing

$$\phi(x) = \frac{1}{\sqrt{2}}[\eta + \varphi_1(x) + i\varphi_2(x)] \quad [16]$$

where φ_1 and φ_2 are real. Then the Lagrangian becomes

$$\begin{aligned} \mathcal{L} = \frac{1}{2}[\dot{\varphi}_1^2 - (\nabla\varphi_1)^2 + \dot{\varphi}_2^2 - (\nabla\varphi_2)^2 - \lambda\eta^2\varphi_1^2 \\ - \lambda\eta\varphi_1(\varphi_1^2 + \varphi_2^2) - \frac{1}{4}\lambda(\varphi_1^2 + \varphi_2^2)^2] \end{aligned} \quad [17]$$

Evidently, the field φ_1 , corresponding to radial oscillations in ϕ , is massive, with mass $\sqrt{\lambda}\eta$. But there is no term in φ_2^2 , so φ_2 is massless.

In the case of spontaneous symmetry breaking of nonabelian symmetries, there may be several Goldstone bosons, one for each broken component of the continuous symmetry. In our theory with symmetry group $G = O(N)$, the possible values of the vacuum expectation value at $T=0$ are $\langle 0_n | \hat{\phi}(0) | 0_n \rangle \approx \eta \mathbf{n}$,

where \mathbf{n} is an arbitrary unit vector. In this case, for given \mathbf{n} , there is an unbroken symmetry subgroup

$$H = \{R \in O(N) : R\mathbf{n} = \mathbf{n}\} = O(N-1) \quad [18]$$

and the number of broken symmetries is

$$\dim G - \dim H = N - 1 \quad [19]$$

Thus, the radial component of ϕ is massive, and there are $N-1$ Goldstone bosons, the $N-1$ transverse components.

Spontaneously Broken Gauge Theories

As we shall see, symmetry breaking in gauge theories is a more problematic concept but, for the moment, these complications are ignored and the present discussion will continue with an approach similar to that used above.

The simplest local gauge symmetry theory is a U(1) Higgs model, a model of a complex scalar field $\phi(x)$ interacting with a gauge potential $A_\mu(x)$, described by the Lagrangian

$$\mathcal{L} = D_\mu \phi^* D^\mu \phi - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - V(|\phi|) \quad [20]$$

where V is a sombrero potential as in [5], while the covariant derivative $D_\mu \phi$ and gauge field $F_{\mu\nu}$ are given by

$$D_\mu \phi = \partial_\mu \phi + ieA_\mu \phi, \quad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \quad [21]$$

The model is invariant under the local U(1) gauge transformations

$$\begin{aligned} \phi(x) &\mapsto \phi(x) e^{i\alpha(x)} \\ A_\mu(x) &\mapsto A_\mu(x) - \frac{1}{e} \partial_\mu \alpha(x) \end{aligned} \quad [22]$$

The Goldstone theorem does not apply to local-symmetry theories. The problem is that to have a Hilbert space containing only physical states one must eliminate the gauge freedom by choosing a gauge condition (e.g., in the U(1) case the Coulomb gauge $\partial_k A^k(x) = 0$, which has the effect of restricting the number of polarization states of photons to two). This necessarily breaks manifest Lorentz invariance, although the theory is, of course, still fully Lorentz invariant. The proof of the theorem fails because the current is no longer local; the long-range Coulomb interaction makes the commutator fall off only like $1/r^2$, so the surface integral no longer vanishes in the infinite-volume limit. (The theorem also fails for nonrelativistic models with long-range forces.)

Again, consider a vacuum state $|0\rangle$ in which $\langle 0|\hat{\phi}|0\rangle = \eta/\sqrt{2}$, and make the same decomposition, [16]. Then, if we set

$$A'_\mu = A_\mu + \frac{1}{e\eta} \partial_\mu \varphi_2 \quad [23]$$

we find that the kinetic term for φ_2 has been absorbed into a mass term $(1/2)e^2\eta^2 A'_\mu A'^\mu$ for the vector field. We have a model with only massive fields: the “Higgs field” φ_1 with mass $\sqrt{\lambda}\eta$ and the gauge field A'_μ with mass $e\eta$. The Goldstone bosons have been “eaten up” by the vector field to provide its longitudinal mode. This is the Higgs mechanism, first noted by Anderson in the context of the photon in a plasma becoming a massive plasmon.

A more elegant way of seeing this is to note that we can always make a gauge transformation to ensure that ϕ is real (at least so long as $\phi \neq 0$; where it is zero, there may be problems). This means that $\phi(x) = (1/\sqrt{2})(\eta + \varphi_1)$; φ_2 disappears altogether, and its kinetic term reduces to $(1/2)e^2 A_\mu A^\mu (\eta + \varphi_1)^2$, which includes the mass term for A_μ as well as cubic and quartic interaction terms.

As before, the discussion can be generalized to nonabelian theories, although there are additional problems to be discussed later. If we have a local symmetry group G that breaks spontaneously to leave an unbroken subgroup H , then the gauge fields associated with H remain massless. Each of the $(\dim G - \dim H)$ complementary fields “eats up” one of the Goldstone bosons, becoming massive in the process. We are left only with other, “radial” components of ϕ , the massive Higgs fields.

Consider, for example, a local SO(3) model, with scalar fields $\phi = (\phi_a)_{a=1,2,3}$ and gauge potentials $A_\mu = (A_{a\mu})$. The infinitesimal gauge transformations are

$$\delta\phi = \delta\omega \times \phi, \quad \delta A_\mu = \delta\omega \times A - \frac{1}{e} \partial_\mu \delta\omega \quad [24]$$

where $\delta\omega$ is the gauge parameter. The Lagrangian is

$$\mathcal{L} = \frac{1}{2} D_\mu \phi \cdot D^\mu \phi - \frac{1}{4} F_{\mu\nu} \cdot F^{\mu\nu} - \frac{1}{8} \lambda (\phi^2 - \eta^2)^2 \quad [25]$$

where the covariant derivative and gauge field are

$$\begin{aligned} D_\mu \phi &= \partial_\mu \phi + eA_\mu \times \phi \\ F_{\mu\nu} &= \partial_\mu A_\nu - \partial_\nu A_\mu + eA_\mu \times A_\nu \end{aligned} \quad [26]$$

If we take $\langle \hat{\phi} \rangle$ in the 3-direction, the fields $A_{1\mu}$ and $A_{2\mu}$ absorb the Goldstone fields ϕ_1, ϕ_2 to become massive. As in the abelian case, we can use the local SO(3) invariance to rotate ϕ everywhere to the 3-direction, and write $\phi = (0, 0, \eta + \varphi_3)$. In this gauge the kinetic term $(1/2)(eA_\mu \times \phi)^2$ gives a mass $e\eta$ to the fields $A_{1\mu}, A_{2\mu}$ while $A_{3\mu}$ remains massless, and the Higgs field φ_3 again has mass $\sqrt{\lambda}\eta$.

Elitzur’s Theorem; the Role of Gauge Fixing

The concept of spontaneous symmetry breaking in the context of a local symmetry requires further discussion, in particular because of Elitzur’s theorem, proved in 1975, which states in essence that “spontaneous breaking of a local symmetry is impossible.” In the light of this theorem, it may seem that a “spontaneously broken gauge theory” is an oxymoron. In fact, it means something rather different, although even that is not unproblematic.

The theorem was proved in the context of lattice gauge theory, where the spatial continuum is replaced by a discrete lattice. The scalar field is then represented by values ϕ_x at each lattice site, and the gauge potential by values $A_{x,\mu}$ on the links of the lattice. This is significant because on the lattice one can use a manifestly gauge-invariant formalism. Expectation values of gauge-invariant physical variables can be found, for example, by a Monte Carlo algorithm that effectively averages over all possible gauges. In this context, it is possible to show that the expectation value of any gauge-noninvariant operator (such as $\hat{\phi}_x$) necessarily vanishes identically.

To be more specific, suppose we incorporate a symmetry-breaking term of the form $-j \cdot \sum_x \phi_x$, and consider the limits $V \rightarrow \infty$ followed by $j \rightarrow 0$. In the global-symmetry case, as we noted earlier, this yields the nonzero result [13]. However, in the case of a local gauge symmetry, one can show rigorously that

$$\lim_{j \rightarrow 0} \lim_{V \rightarrow \infty} \langle \hat{\phi}_x \rangle_{jn} = 0 \tag{27}$$

The essential reason for this is that we can make a gauge transformation in the neighborhood of the point x to make ϕ_x have any value we like without changing the energy by more than a very small amount that goes to zero as $j \rightarrow 0$. Within this manifestly gauge-invariant formalism, it is clear that the expectation value of a gauge-noninvariant operator such as $\hat{\phi}$ is not an appropriate order parameter. One must instead look for a gauge-invariant order parameter.

It is important to note, however, that this result applies only in the context of a manifestly gauge-invariant formalism. But, in general, gauge theories cannot be quantized in a manifestly gauge-invariant way. In a path-integral formalism, the action functional, which appears in the exponent, is constant along the orbits of the gauge-group action. Consequently, the integral contains an infinite factor, the volume of the (infinite-dimensional) gauge group. There are corresponding divergences

in the perturbation series. As is well known, this problem can be dealt with by introducing a gauge-fixing term, which explicitly breaks the gauge symmetry, and renders Elitzur’s theorem inapplicable. But this procedure leaves a global symmetry unbroken, and it is in fact that global symmetry that is broken spontaneously.

One example is the Landau–Ginzburg model of a superconductor, which is essentially just the non-relativistic limit of the abelian Higgs model, although there is one significant difference: here the field $\hat{\phi}$ annihilates a Cooper pair, a bound pair of electrons with equal and opposite momenta and spins, so e above is replaced by the charge $2e$ of a Cooper pair. The appearance of a condensate of Cooper pairs in the low-temperature superconducting phase corresponds to a state in which $\langle \hat{\phi} \rangle$ is nonzero. This would not be possible without fixing a gauge. In the nonrelativistic context, the obvious gauge to choose is the Coulomb gauge, defined by the condition $\partial_k A^k = 0$. This gauge-fixing condition breaks the local symmetry explicitly, but it leaves unbroken the global symmetry $\phi(x) \rightarrow \phi(x)e^{i\alpha}$ with constant α . It is that global symmetry that is spontaneously broken when $\langle \hat{\phi} \rangle \neq 0$.

For a model with nonabelian local symmetry the standard procedure used to derive a perturbation expansion is that of Faddeev and Popov. Consider, for example, the $SO(3)$ gauge theory discussed in the preceding section. To fix the gauge, we can choose a set of functions $F = (F_a)$ of the fields, and introduce into the path integral a gauge-fixing term of the form

$$\mathcal{L}_{\text{gf}} = -\frac{1}{2\xi} F^2 \tag{28}$$

where ξ is an arbitrary real constant. However, to ensure that this does not bias the integral, so that the gauge-fixed theory is at least formally equivalent to the original gauge-invariant theory, one must also include the determinant of the Jacobian matrix

$$J_{ab}(x,y) = \frac{\delta F_a(x)}{\delta \omega_b(y)} \tag{29}$$

The easiest way to do this is to introduce Faddeev–Popov ghost fields \bar{C}, C , which are scalar Grassmann variables, and an appropriate term in the Lagrangian

$$\mathcal{L}_{\text{FP}} = \bar{C} \cdot J \cdot C \tag{30}$$

For the $SO(3)$ model, a convenient choice of gauge is the R_ξ gauge defined by

$$F = \partial_\mu A^\mu - \xi e \eta n \times \phi \tag{31}$$

where \mathbf{n} is an arbitrarily chosen unit vector. It is clear that the full Lagrangian $\mathcal{L} + \mathcal{L}_{\text{gf}} + \mathcal{L}_{\text{FP}}$ is no longer invariant under the full $\text{SO}(3)$ gauge group, although there is a residual $\text{U}(1)$ gauge invariance corresponding to rotations about \mathbf{n} . In this gauge, the arbitrary choice of \mathbf{n} means that the global $\text{SO}(3)$ symmetry is also broken. However, for other choices, such as the Lorentz gauge $F = \partial_\mu A^\mu$ or axial gauge $F = A_3$, the Lagrangian is invariant under global $\text{SO}(3)$ rotations of all the fields. This global symmetry is then spontaneously broken, with $\hat{\phi}$ acquiring as before a nonzero expectation value of the form $\langle \hat{\phi}(x) \rangle = \eta \mathbf{n}$.

It is interesting to look again at the particle content of this model. By setting $\phi(x) = \eta \mathbf{n} + \varphi(x)$ with $\mathbf{n} = (0, 0, 1)$, one finds that in the quadratic part of the Lagrangian, the cross-terms between A_μ and φ combine to form a total divergence which can be dropped. As before, φ_3 is the Higgs field, with $m^2 = \lambda \eta^2$, $A_{3\mu}$ is the massless gauge field corresponding to the unbroken gauge symmetry, and the three transverse components of $A_{1\mu}$ and $A_{2\mu}$ represent the massive vector fields, with $m^2 = e^2 \eta^2$. There are, however, also unphysical fields with ξ -dependent masses: $\varphi_{1,2}$, $C_{1,2}$, $\bar{C}_{1,2}$, and the longitudinal components $\partial_\mu A_{1,2}^\mu$ all have $m^2 = \xi e^2 \eta^2$. We can now compute the effective potential $V_{\text{eff}}(T, \varphi)$. One point that should be noted in performing this calculation is that the ghost fields C, \bar{C} contribute negatively. Obviously, V_{eff} , being ξ -dependent, is not itself physically meaningful. Nevertheless, it can be shown that the stationary points of V_{eff} are physical, and correspond to the possible equilibrium states of the theory. Moreover, the extremal values of V_{eff} are independent of ξ and give correctly the thermodynamic potential in the corresponding equilibrium states. The negative contributions from the ghost fields to N_* and M_*^2 ensure that the ξ dependence cancels out, and we find as expected $N_* = 9$ and $M_*^2 = (\lambda + 6e^2)\eta^2$.

Phase Transitions and Crossovers

Our discussion so far has for the most part been restricted to a semiclassical or mean-field approximation. It is important to bear in mind, however, that this approximation does not suffice to determine whether a phase transition (where the thermodynamic free energy is nonanalytic) exists, or what its nature is. Determining the detailed characteristics of phase transitions requires other methods, such as the renormalization group or lattice simulations. In many cases, it is far from trivial to establish the order of the transition, or even whether a true phase transition actually exists.

Gauge theories pose particular problems because of the infrared divergences in the thermal field theory at high temperature, and because in asymptotically free nonabelian theories the coupling becomes large at very low energy. Even when they appear to exhibit spontaneous symmetry breaking, they do not necessarily undergo a true phase transition. Lattice gauge theory calculations have led to the conclusion that in nonabelian gauge theories with the Higgs field in the fundamental representation, there are values of the coupling constants for which there is no phase transition, only a rapid but smooth crossover from one type of behavior to another, so that the high- and low-temperature phases are analytically connected. If the coupling constant is small, there is a first-order phase transition, and for moderate values the theory exhibits a very rapid crossover that looks quite similar to a symmetry-breaking phase transition. Nevertheless, the analytic connection between the two phases implies that there cannot exist an order parameter that is strictly zero above the transition and nonzero below it.

In particular, it appears that for physical values of the Higgs mass, the electroweak theory does not undergo in fact undergo a true phase transition. It is somewhat ironic that the most famous example of a spontaneously broken gauge theory probably does not, strictly speaking, exhibit a symmetry-breaking phase transition!

Conclusions

We have discussed the main features of spontaneous symmetry breaking in both the global- and local-symmetry cases, especially the appearance of Goldstone bosons when a continuous global symmetry breaks, and their elimination in the local-symmetry case by the Higgs mechanism, as well as the problems attaching to the concept of spontaneous symmetry breaking in gauge theories.

See also: Abelian Higgs Vortices; Effective Field Theories; Electroweak Theory; Finite Group Symmetry Breaking; Lattice Gauge Theory; Noncommutative Geometry and the Standard Model; Phase Transitions in Continuous Systems; Quantum Central Limit Theorems; Quantum Spin Systems; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Topological Defects and their Homotopy Classification.

Further Reading

Anderson PW (1963) Plasmons, gauge invariance and mass. *Physical Review* 130: 439–442.

- Coleman S (1985) *Aspects of Symmetry*, ch. 5. Cambridge: Cambridge University Press.
- Elitzur S (1975) Impossibility of spontaneously breaking local symmetries. *Physical Review D* 12: 3978–3982.
- Englert F and Brout R (1964) Broken symmetry and the mass of gauge vector bosons. *Physical Review Letters* 13: 321–323.
- Fradkin E and Shenker SH (1979) Phase diagrams of lattice gauge theories with Higgs fields. *Physical Review D* 19: 3682–3697.
- Goldstone J, Salam A, and Weinberg S (1962) Broken symmetries. *Physical Review* 127: 965–970.
- Guralnik GS, Hagen CR, and Kibble TWB (1964) Global conservation laws and massless particles. *Physical Review Letters* 13: 585–587.
- Higgs PW (1964) Broken symmetries, massless particles and gauge fields. *Physics Letters* 12: 132–133.
- Kibble TWB (1967) Symmetry-breaking in non-abelian gauge theories. *Physical Review* 155: 1554–1561.
- Weinberg S (1996) *The Quantum Theory of Fields, Vol. II. Modern Applications*, chs. 19 and 21. Cambridge: Cambridge University Press.

Symmetry Classes in Random Matrix Theory

M R Zirnbauer, Universität Köln, Köln, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

A classification of random matrix ensembles by symmetries was first established by Dyson, in an influential 1962 paper with the title “the threefold way: algebraic structure of symmetry groups and ensembles in quantum mechanics.” Dyson’s threefold way has since become fundamental to various areas of theoretical physics, including the statistical theory of complex many-body systems, mesoscopic physics, disordered electron systems, and the field of quantum chaos.

Over the last decade, a number of random matrix ensembles beyond Dyson’s classification have come to the fore in physics and mathematics. On the physics side, these emerged from work on the low-energy Dirac spectrum of quantum chromodynamics (QCD) and from the mesoscopic physics of low-energy quasiparticles in disordered superconductors. In the mathematical research area of number theory, the study of statistical correlations in the values of Riemann zeta and similar functions has prompted some of the same generalizations.

In this article, Dyson’s fundamental result will be reviewed from a modern perspective, and the recent extension of Dyson’s threefold way will be motivated and described. In particular, it will be explained why symmetry classes are associated with large families of symmetric spaces.

The Framework

Random matrices have their physical origin in the quantum world, more precisely in the statistical theory of strongly interacting many-body systems such as atomic nuclei. Although random matrix theory is nowadays understood to be of relevance to

numerous areas of physics – see Random Matrix Theory in Physics – quantum mechanics is still where many of its applications lie. Quantum mechanics also provides a natural framework in which to classify random matrix ensembles.

Following Dyson, the mathematical setting for classification consists of two pieces of data:

- A finite-dimensional complex vector space V with a Hermitian scalar product $\langle \cdot, \cdot \rangle$, called a “unitary structure” for short. (In physics applications, V will usually be the truncated Hilbert space of a family of quantum Hamiltonian systems.)
- On V there acts a group G of unitary and antiunitary operators (the joint symmetry group of the multiparameter family of quantum systems).

Given this setup, one is interested in the linear space of self-adjoint operators on V – the Hamiltonians H – with the property that they commute with the G -action. Such a space is reducible in general, that is, the matrix of H decomposes into blocks. The goal of classification is to list all of the irreducible blocks that occur.

Symmetry Groups

Basic to classification is the notion of a symmetry group in quantum Hamiltonian systems, a notion that will now be explained.

In classical mechanics, the symmetry group G_0 of a Hamiltonian system is understood to be the group of canonical transformations that commute with the phase flow of the system. An important example is the rotation group for systems in a central field.

In passing from classical to quantum mechanics, one replaces the classical phase space by a quantum-mechanical Hilbert space V and assigns to the symmetry group G_0 a (projective) representation by unitary \mathbb{C} -linear operators on V . Besides the one-parameter continuous subgroups, whose significance is highlighted by Noether’s theorem, the components of G_0 not connected with the identity play an

important role. A prominent example is provided by the operator for space reflection. Its eigenspaces are the subspaces of states with positive and negative parity; these reduce the matrix of any reflection-invariant Hamiltonian to two blocks.

Not all symmetries of a quantum-mechanical system are of the canonical, unitary kind: the prime counterexample is the operation of inverting the time direction, called time reversal for short. In classical mechanics, this operation reverses the sign of the symplectic structure of phase space; in quantum mechanics, its algebraic properties reflect the fact that inverting the time direction, $t \mapsto -t$, amounts to sending $i = \sqrt{-1}$ to $-i$. Indeed, time t enters in the Dirac, Pauli, or Schrödinger equation as $i\hbar d/dt$. Therefore, time reversal is represented in the quantum theory by an antiunitary operator T , which is to say that T is complex antilinear:

$$T(zv) = \bar{z}Tv \quad (z \in \mathbb{C}, v \in V)$$

and preserves the Hermitian scalar product or unitary structure up to complex conjugation:

$$\langle Tv_1, Tv_2 \rangle = \overline{\langle v_1, v_2 \rangle} = \langle v_2, v_1 \rangle$$

Another operation of this kind is charge conjugation in relativistic theories such as the Dirac equation.

By the symmetry group G of a quantum-mechanical system with Hamiltonian H , one then means the group of all unitary and antiunitary transformations g of V that leave the Hamiltonian invariant: $gHg^{-1} = H$. We denote the unitary subgroup of G by G_0 , and the set of antiunitary operators in G by G_1 (not a group). If V carries extra structure, as will be the case for some extensions of Dyson's basic scheme, the action of G on V has to be compatible with that structure.

The set G_1 may be empty. When it is not, the composition of any two elements of G_1 is unitary, so every $g \in G_1$ can be obtained from a fixed element of G_1 , say T , by right multiplication with some $U \in G_0$: $g = TU$. In other words, when G_1 is nonempty the coset space G/G_0 consists of exactly two elements, G_0 and $T \cdot G_0 = G_1$. We shall assume that T represents some inversion symmetry such as time reversal or charge conjugation. T must then be a (projective) involution, that is, $T^2 = z \times \text{Id}$ with z a complex number of unit modulus, so that conjugation by T^2 is the identity operation. Since T is complex antilinear, the associative law $T^2 \cdot T = T \cdot T^2$ forces z to be real, and hence $T^2 = \pm \text{Id}$.

Finding the total symmetry group of a Hamiltonian system need not always be straightforward, but this complication will not be an issue here: we take the symmetry group G and its action on the Hilbert space V as fundamental and given, and then ask

what are the corresponding symmetry classes, meaning the irreducible spaces of Hamiltonians on V that commute with G .

For technical reasons, we assume the group G_0 to be compact; this is an assumption that covers most (if not all) of the cases of interest in physics. The noncompact group of space translations can be incorporated, if necessary, by wrapping the system around a torus, whereby translations are turned into compact torus rotations.

While the primary objects to classify are the spaces of Hamiltonians H , we shall focus for convenience on the spaces of time evolutions $U_t = e^{-itH/\hbar}$ instead. This change of focus results in no loss, as the Hamiltonians can always be retrieved by linearizing in t at $t = 0$.

Symmetric Spaces

We appropriate a few basic facts from the theory of symmetric spaces.

Let M be a connected m -dimensional Riemannian manifold and p a point of M . In some open subset N_p of a neighborhood of p there exists a map $s_p: N_p \rightarrow N_p$, the geodesic inversion with respect to p , which sends a point $x \in N_p$ with normal coordinates (x_1, \dots, x_m) to the point with normal coordinates $(-x_1, \dots, -x_m)$. The Riemannian manifold M is called locally symmetric if the geodesic inversion is an isometry, and is called globally symmetric if s_p extends to an isometry $s_p: M \rightarrow M$, for all $p \in M$. A globally symmetric Riemannian manifold is called a symmetric space for short.

The Riemann curvature tensor of a symmetric space is covariantly constant, which leads one to distinguish between three cases: the scalar curvature can be positive, zero, or negative, and the symmetric space is said to be of compact type, Euclidean type, or noncompact type, respectively. (In mesoscopic physics, each type plays a role: the first provides us with the scattering matrices and time evolutions, the second with the Hamiltonians, and the third with the transfer matrices.) The focus in the current article will be on compact type, as it is this type that houses the unitary time evolution operators of quantum mechanics. The compact symmetric spaces are subdivided into two major subtypes, both of which occur naturally in the present context, as follows.

Type II

Consider first the case where the antiunitary component G_1 of the symmetry group is empty, so the data are (V, G) with $G = G_0$. Let $\mathcal{L}(V)$ denote the group of all complex linear transformations that

leave the structure of the vector space V invariant. Thus, $\mathcal{U}(V)$ is a group of unitary transformations if V carries no more than the usual Hermitian scalar product; and is some subgroup of the unitary group if V does have extra structure (as is the case for the Nambu space of quasiparticle excitations in a superconductor). The symmetry group G_0 , by acting on V and preserving its structure, is contained as a subgroup in $\mathcal{U}(V)$.

Let now H be any Hamiltonian with the prescribed symmetries. Then the time evolution $t \mapsto U_t = e^{-itH/\hbar}$ generated by H is a one-parameter subgroup of $\mathcal{U}(V)$ which commutes with the G_0 -action. The total set of transformations U_t that arise in this way is called the (connected part of the) “centralizer” of G_0 in $\mathcal{U}(V)$, and is denoted by Z . This is the “good” set of unitary time evolutions – the set compatible with the given symmetries of an ensemble of quantum systems.

The centralizer Z is obviously a group: if U and U' belong to Z , then so do their inverses and their product. What can one say about the structure of the group Z ?

Since G_0 is compact by assumption, its group action on V is completely reducible and V is guaranteed to have an orthogonal decomposition

$$V = \bigoplus_{\lambda} V_{\lambda}$$

where the sum runs over isomorphism classes of irreducible G_0 -representations λ , and the vector spaces V_{λ} are called the G_0 -isotypic components of V . For example, if G_0 is the rotation group SO_3 , the G_0 -isotypic component V_{λ} of V is the subspace spanned by all the states with total angular momentum λ .

Consider now any $U \in Z$. Since U commutes with the G_0 -action, it does not connect different G_0 -isotypic components. (Indeed, in the example of SO_3 -invariant dynamics, angular momentum is conserved and transitions between different angular momentum sectors are forbidden.) Thus, every G_0 -isotypic component V_{λ} is an invariant subspace for the action of Z on V , and Z decomposes as $Z = \prod_{\lambda} Z_{\lambda}$ with blocks $Z_{\lambda} = Z|_{V_{\lambda}}$.

To say more, fix a standard irreducible G_0 -module R_{λ} of isomorphism class λ and consider

$$L_{\lambda} := \text{Hom}_{G_0}(R_{\lambda}, V_{\lambda})$$

the linear space of \mathbb{C} -linear maps $l: R_{\lambda} \rightarrow V_{\lambda}$ that intertwine the G_0 -actions on R_{λ} and V_{λ} . An element of L_{λ} is called a G_0 -equivariant homomorphism. By Schur’s lemma, $L_{\lambda} \cong \mathbb{C}$ if V_{λ} is G_0 -irreducible. More generally, $\dim L_{\lambda} = m_{\lambda}$ counts the multiplicity of occurrence of R_{λ} in V_{λ} ; for example, in the case of

$G_0 = SO_3$ we take R_{λ} to be the standard irreducible module of dimension $2\lambda + 1$; and m_{λ} then is the number of times a multiplet of states with total angular momentum λ occurs in V_{λ} .

The natural mapping $L_{\lambda} \otimes R_{\lambda} \rightarrow V_{\lambda}$ by $l \otimes r \mapsto l(r)$ is an isomorphism,

$$V_{\lambda} \cong L_{\lambda} \otimes R_{\lambda}$$

and using it we can transfer the entire discussion from V_{λ} to $L_{\lambda} \otimes R_{\lambda}$. The group G_0 acts trivially on $L_{\lambda} \cong \mathbb{C}^{m_{\lambda}}$ and irreducibly on R_{λ} . Therefore, the component Z_{λ} of the centralizer Z is the unitary group

$$Z_{\lambda} \cong U(L_{\lambda}) \cong U_{m_{\lambda}}$$

if V is a unitary vector space with no extra structure. In the presence of extra structure (which, by compatibility with the G_0 -action, restricts to every subspace V_{λ}), the factor Z_{λ} is some subgroup of $U_{m_{\lambda}}$. In all cases, Z is a direct product of connected compact Lie groups Z_{λ} .

To make the connection with symmetric spaces, write $M := Z_{\lambda}$. Since M is a group, the operation of taking the inverse, $U \mapsto U^{-1}$, makes sense for all $U \in M$. Moreover, being a compact Lie group, the manifold M admits a left- and right-invariant Riemannian structure in which the inversion $U \mapsto U^{-1}$ is an isometry. By translation, one gets an isometry $s_{U_1}: U \mapsto U_1 U^{-1} U_1$ for every $U_1 \in M$. All of these maps s_{U_1} are globally defined, and the restriction of s_{U_1} to some neighborhood of U_1 coincides with the geodesic inversion with respect to U_1 . Thus, M is a symmetric space by the definition given above. Symmetric spaces of this kind are called type II.

Type I

Consider next the case of $G_1 \neq \emptyset$, where some antiunitary symmetry T is present. As before, let Z be the connected component of the centralizer of G_0 in $\mathcal{U}(V)$. Conjugation by T ,

$$U \mapsto \tau(U) := TUT^{-1}$$

is an automorphism of $\mathcal{U}(V)$ and, owing to $T^2 = \pm \text{Id}$, τ is involutive. Because $G_0 \subset G$ is a normal subgroup, τ restricts to an involutive automorphism (still denoted by τ) of Z . Now recall that T is complex antilinear and the good Hamiltonians are subject to $THT^{-1} = H$. The good time evolutions $U_t = e^{-itH/\hbar}$ clearly satisfy $\tau(U_t) = U_{-t} = U_t^{-1}$. Thus, the good set to consider is $\mathcal{U} := \{U \in Z \mid U = \tau(U)^{-1}\}$. The set \mathcal{U} is a manifold, but in general is not a Lie group.

Further details depend on what τ does with the factorization $Z = \prod_{\lambda} Z_{\lambda}$. If V_{λ} is a G_0 -isotypic

component of V , then so is TV_λ , since T normalizes G_0 . Thus, either $V_\lambda \cap TV_\lambda = 0$, or $TV_\lambda = V_\lambda$. In the former case, the involutive automorphism τ just relates $U \in Z_\lambda$ with $\tau(U) \in Z_{TV_\lambda}$, whence no intrinsic constraint on Z_λ results, and the time evolutions $(U, \tau(U)^{-1}) \in Z_\lambda \times Z_{TV_\lambda}$ constitute a type-II symmetric space, as before.

A novel situation occurs when $TV_\lambda = V_\lambda$, in which case τ restricts to an automorphism of Z_λ . Let therefore $TV_\lambda = V_\lambda$, put $K \equiv Z_\lambda$ for short, and consider

$$M := \{U \in K | U = \tau(U)^{-1}\}$$

Note that if two elements p, p_0 of K are in M , then so is the product $p_0 p^{-1} p_0$. The group K acts on $M \subset K$ by

$$k \cdot U = k U \tau(k)^{-1} \quad (k \in K)$$

and this group action is transitive, that is, every $U \in M$ can be written as $U = k \tau(k)^{-1}$ with some $k \in K$. (Finding k for a given U is like taking a square root, which is possible since $\exp : \text{Lie } K \rightarrow K$ is surjective.) There exists such a K -invariant Riemannian structure for M that for all $p_0 \in M$ the mapping $s_{p_0} : M \rightarrow M$ defined by

$$s_{p_0}(p) = p_0 p^{-1} p_0$$

is the geodesic inversion with respect to $p_0 \in M$. Thus, in this natural geometry M is a globally symmetric Riemannian manifold and hence a symmetric space. The present kind of symmetric space is called type I. If K_τ is the set of fixed points of τ in K , the symmetric space M is analytically diffeomorphic to the coset space K/K_τ by

$$K/K_\tau \rightarrow M \subset K, \quad UK_\tau \mapsto U\tau(U)^{-1}$$

which we call the “Cartan embedding” of K/K_τ into K .

In summary, the solution to the problem of finding the set of unitary time evolution operators that are compatible with a given symmetry group G and structure of Hilbert space V is always a symmetric space. This is a valuable insight, as symmetric spaces are rigid objects and have been completely classified by Cartan.

If the dimension of V is kept variable, the irreducible symmetric spaces that occur belong to one of the large families listed in Table 1.

Dyson’s Threefold Way

Recall the goal: given a Hilbert space V and a symmetry group G acting on it, one wants to classify

Table 1 The large families of symmetric spaces. The form of H in the header applies to the last seven families

Family	Symmetric space	Form of $H = \begin{pmatrix} W & Z \\ Z^\dagger & -W \end{pmatrix}$
A	U_N	Complex Hermitian
AI	U_N/O_N	Real symmetric
AII	U_{2N}/USp_{2N}	Quaternion self-adjoint
C	USp_{2N}	Z complex symmetric, $W = W^i$
CI	USp_{2N}/U_N	Z complex symmetric, $W = 0$
D	SO_{2N}	Z complex skew, $W = W^i$
DIII	SO_{2N}/U_N	Z complex skew, $W = 0$
AIII	$U_{p+q}/U_p \times U_q$	Z complex $p \times q$, $W = 0$
BDI	$SO_{p+q}/SO_p \times SO_q$	Z real $p \times q$, $W = 0$
CII	$USp_{2p+2q}/USp_{2p} \times USp_{2q}$	Z quaternion $2p \times 2q$, $W = 0$

the (irreducible) spaces of time evolution operators U that are “compatible” with G , meaning

$$U = g_0 U g_0^{-1} = g_1 U^{-1} g_1^{-1} \quad (\text{for all } g_\sigma \in G_\sigma)$$

As we have seen, the spaces that arise in this way are symmetric spaces of type I or II depending on the nature of the time reversal (or other antiunitary symmetry) T .

An even stronger statement can be made when more information about the Hilbert space V is specified. In Dyson’s classification, the Hermitian scalar product of V is assumed to be the only invariant structure that exists on V . With that assumption, only three large families of symmetric spaces arise; these correspond to what we call the “Wigner–Dyson symmetry classes.”

Class A

Recall that in Dyson’s case, the connected part of the centralizer of G_0 in $\mathscr{U}(V)$ is a direct product of unitary groups, each factor being associated with one G_0 -isotypic component V_λ of V . The type-II situation occurs when the set G_1 of antiunitary symmetries is either empty or else exchanges different V_λ . In both cases, the set of good time evolution operators restricted to one G_0 -isotypic component V_λ is a unitary group U_{m_λ} , with m_λ being the multiplicity of the irreducible G_0 -representation λ in V_λ .

The unitary groups $U_{N=m_\lambda}$ or, to be precise, their simple parts SU_N , are called type-II symmetric spaces of the A family or A series – hence the name class A. The Hamiltonians H , the generators of time evolutions $U_t = e^{-itH/\hbar}$, in this class are represented by

complex Hermitian $N \times N$ matrices. By putting a U_N -invariant Gaussian probability measure

$$\exp(-\text{tr}H^2/2\sigma^2) dH \quad (\sigma \in \mathbb{R})$$

on that space, one gets what is called the GUE – the Gaussian unitary ensemble – which defines the Wigner–Dyson universality class of unitary symmetry.

Classes AI and AII

Consider next the case $G_1 \neq \emptyset$, with antiunitary generator T . Let $V_\lambda = TV_\lambda$ be any G_0 -isotypic component of V invariant under T (the type-I situation). The mapping $U \mapsto TUT^{-1} = \tau(U)$ then is an automorphism of the groups $U(V_\lambda), G_0$ and $K = Z_\lambda \cong U_{m_\lambda}$. If K_τ is the subgroup of fixed points of τ in K , the space of good time evolutions can be identified with the symmetric space K/K_τ by the Cartan embedding. Our task is to determine K_τ .

To simplify the notation let us write $V_\lambda \equiv V, R_\lambda \equiv R$, and $L_\lambda \equiv L$. We now ask what happens with $T: V \rightarrow V$ in the process of transfer to $L \otimes R \cong V$. The answer, so we claim, is that T transfers to a *pure* tensor made from antiunitary maps $\alpha: L \rightarrow L$ and $\beta: R \rightarrow R$,

$$T = \alpha \otimes \beta$$

To prove this claim, let C be the antilinear map from V to the *dual* vector space V^* by $v \mapsto \langle v, \cdot \rangle$. Because the elements of G_0 are represented by unitaries, the \mathbb{C} -linear operator $CT: V \rightarrow V^*$ intertwines G_0 -actions:

$$CTa(g) = g^{-1*}CT \quad (g \in G_0)$$

where a is the automorphism $a(g) = T^{-1}gT$. From the irreducibility of R it follows that the space of intertwiners $R \rightarrow R^*$ is one dimensional here (Schur's lemma). Therefore, $CT: L \otimes R \rightarrow L^* \otimes R^*$ must be a pure tensor (as opposed to a sum of such tensors), and since C is clearly a pure tensor, so is T . This completes the proof.

By the involutive property $T^2 = \varepsilon_T \text{Id}_V$ ($\varepsilon_T = \pm 1$), the two antiunitary factors of $T = \alpha \otimes \beta$ cannot but square to $\alpha^2 = \varepsilon_\alpha \text{Id}_L$ and $\beta^2 = \varepsilon_\beta \text{Id}_R$ where $\varepsilon_\alpha, \varepsilon_\beta = \pm 1$ are related by $\varepsilon_\alpha \varepsilon_\beta = \varepsilon_T$. The factor α determines a nondegenerate complex bilinear form $Q: L \times L \rightarrow \mathbb{C}$ by

$$Q(l_1, l_2) = \langle \alpha l_1, l_2 \rangle_L \quad (l_1, l_2 \in L)$$

Since α is antiunitary one has the exchange symmetry

$$Q(l_1, l_2) = \overline{\langle \alpha^2 l_1, \alpha l_2 \rangle_L} = \varepsilon_\alpha Q(l_2, l_1)$$

Thus, the complex bilinear form (or pairing) Q is symmetric for $\varepsilon_\alpha = +1$ and alternating for $\varepsilon_\alpha = -1$.

Knowing the sign of $\varepsilon_\alpha = \pm 1$ we know the group K_τ . Indeed, an element $k \in K_\tau$ commutes with T and after transfer from V to L still commutes with α . But since K_τ is a subgroup of $K = U_{m_\lambda}$, this means that $k \in K_\tau$ preserves Q . In the case of $\varepsilon_\alpha = +1$, what is preserved is a symmetric pairing, and therefore $K_\tau \cong O_{m_\lambda}$. For $\varepsilon_\alpha = -1$, the multiplicity m_λ must be even and K_τ preserves an alternating pairing (or symplectic structure); in that case $K_\tau \cong \text{USp}_{m_\lambda}$, the unitary symplectic group.

Thus, there is a dichotomy for the sets of good time evolutions $M \cong K/K_\tau$:

$$\text{Class AI: } K/K_\tau \cong U_N/O_N \quad (N = m_\lambda)$$

$$\text{Class AII: } K/K_\tau \cong U_{2N}/\text{USp}_{2N} \quad (2N = m_\lambda)$$

Again we are referring to symmetric spaces by the names they – or rather their simple parts SU_N/SO_N and SU_{2N}/USp_{2N} – have in the Cartan classification.

In general, there is no immediate means of predicting the parity ε_α , and one has no choice but to go through the steps of constructing α . If $\beta: R \rightarrow R$ happens to be G_0 -invariant, however, the situation simplifies. In that case β determines a G_0 -invariant pairing $R \times R \rightarrow \mathbb{C}$ (in the same way as α determines $Q: L \times L \rightarrow \mathbb{C}$ above). On general grounds, an irreducible G_0 -representation space admits at most one such pairing. If that pairing is symmetric, then, as we have seen, $\varepsilon_\beta = 1$; if it is alternating, then $\varepsilon_\beta = -1$. The parity ε_α is given by $\varepsilon_\alpha \varepsilon_\beta = \varepsilon_T$.

Example Consider any physical system with spin-rotation symmetry ($G_0 = \text{SU}_2$) and time-reversal symmetry. The physical operation of time reversal, T , commutes with spin rotations and, hence, here is a case where the factor β in $T = \alpha \otimes \beta$ is G_0 -invariant. On fundamental physics grounds one has $T^2 = (-1)^{2S}$ on states with spin S . The spin- S representation of SU_2 is known to carry an invariant pairing which is symmetric or skew depending on whether the integer $2S$ is even or odd. Therefore, $\varepsilon_T = \varepsilon_\beta$ and $\varepsilon_\alpha = +1$ in all cases.

Thus, T -invariant systems with no symmetries other than energy and spin invariably are class AI. By breaking spin-rotation symmetry ($G_0 = \{\text{Id}\}$, $\varepsilon_\beta = +1$) while maintaining T -symmetry for states with half-integer spin (say single electrons, which carry spin $S = 1/2$), one gets $\varepsilon_\alpha = -1$, thereby realizing class AII.

The Hamiltonians By passing to the tangent space of K/K_τ at unity one obtains Hermitian matrices with entries that are real numbers (class AI) or real quaternions (class AII). When K_τ -invariant Gaussian

probability measures (called GOE resp., GSE) are put on these spaces, one gets the Wigner–Dyson universality classes of orthogonal resp., symplectic symmetry. In mesoscopic physics, these are realized in disordered metals with time-reversal invariance (absence of magnetic fields and magnetic impurities). Spin-rotation symmetry is broken by strong spin–orbit scatterers such as gold impurities.

Warning

The word “symmetry class” is not synonymous with “universality class.” Indeed, inside a symmetry class many different types of physical behavior are possible. For example, random matrix models for disordered metallic grains with time-reversal symmetry belong to the symmetry class of the example above (class AI), and so do Anderson tight-binding models with real hopping. The former are known to exhibit energy level statistics of universal GOE type, whereas the latter have localized eigenfunctions and hence level statistics which is expected to approach the Poisson limit when the system size goes to infinity.

Disordered Superconductors

When Dirac first wrote down his famous equation in 1928, he assumed that he was writing an equation for the wave function of the electron. Later, because of the instability caused by negative-energy solutions, the Dirac equation was reinterpreted (via second quantization) as an equation for the fermionic field operators of a quantum field theory. A similar change of viewpoint is carried out in reverse in the Hartree–Fock–Bogoliubov mean-field description of quasiparticle excitations in superconductors. There, one starts from the equations of motion for linear superpositions of the electron creation and annihilation operators, and reinterprets them as a unitary quantum dynamics for what might be called the quasiparticle “wave function.”

In both cases – the Dirac equation and the quasiparticle dynamics of a superconductor – there enters a structure not present in the standard quantum mechanics underlying Dyson’s classification: the field operators for fermionic particles are subject to a set of relations called the “canonical anticommutation relations,” and these are preserved by the quantum dynamics. Therefore, whenever second quantization is undone (assuming it can be undone) to return from field operators to wave functions, the wave-function dynamics is required to preserve some extra structure. This puts a linear constraint on the good Hamiltonians H . For our

purposes, the best viewpoint to take is to attribute the extra invariant structure to the Hilbert space V , thereby turning it into a Nambu space.

Nambu Space

Adopting the standard physics conventions of second quantization, consider some set of single-particle creation and annihilation operators c_i^\dagger and c_i , where $i = 1, 2, \dots$ labels an orthonormal system of single-particle states. Such operators are subject to the canonical anticommutation relations (CARs)

$$\begin{aligned} c_i^\dagger c_j + c_j c_i^\dagger &= \delta_{ij} \\ c_i^\dagger c_j^\dagger + c_j^\dagger c_i^\dagger &= 0 = c_i c_j + c_j c_i \end{aligned} \quad [1]$$

When written in terms of $c_j + c_j^\dagger$ and $i(c_j - c_j^\dagger)$, these become the standard defining relations of a Clifford algebra over \mathbb{R} . Field operators are linear combinations $\psi = \sum_i (u_i c_i + f_i c_i^\dagger)$ with complex coefficients u_i and f_i .

Now take H to be some Hamiltonian which is quadratic in the creation and annihilation operators:

$$H = \sum_{i,j} W_{ij} c_i^\dagger c_j + \frac{1}{2} \sum_{i,j} \left(Z_{ij} c_i^\dagger c_j^\dagger + \bar{Z}_{ij} c_j c_i \right)$$

and let H act on field operators ψ by the commutator: $H \cdot \psi \equiv [H, \psi]$. The time evolution of ψ is then determined by the Heisenberg equation of motion

$$-i\hbar \frac{d\psi}{dt} = H \cdot \psi \quad [2]$$

which integrates to $\psi(t) = e^{itH/\hbar} \cdot \psi(0)$, and is easily verified to preserve the CARs [1].

The dynamical equation [2] is equivalent to a system of linear differential equations for the amplitudes u_i and f_i . If these are assembled into vectors, and the W_{ij} and Z_{ij} into matrices, eqn [2] becomes

$$-i\hbar \frac{d}{dt} \begin{pmatrix} f \\ u \end{pmatrix} = \begin{pmatrix} W & Z \\ Z^\dagger & -\bar{W} \end{pmatrix} \begin{pmatrix} f \\ u \end{pmatrix}$$

The Hamiltonian matrix on the right-hand side has some special properties due to $Z_{ij} = -Z_{ji}$ (from $c_i c_j = -c_j c_i$) and $W_{ij} = \bar{W}_{ji}$ (from H being self-adjoint as an operator in Fock space). To keep track of these properties while imposing some unitary and antiunitary symmetries, it is best to put everything in invariant form.

So, let U be the unitary vector space of annihilation operators $u = \sum_i u_i c_i$, and view the creation operators $f = \sum_i f_i c_i^\dagger$ as lying in the dual vector space U^* . The field operators $\psi = u + f$ then are elements of the direct sum $U \oplus U^* =: V$, called “Nambu

space.” On V there exists a canonical unitary structure expressed by

$$\langle \psi, \tilde{\psi} \rangle = \sum_i (\bar{u}_i \tilde{u}_i + \bar{f}_i \tilde{f}_i)$$

A second canonical structure on $V = U \oplus U^*$ is given by the symmetric complex bilinear form

$$\{ \psi, \tilde{\psi} \} = \sum_i (\tilde{f}_i u_i + f_i \tilde{u}_i) = \tilde{f}(u) + f(\tilde{u})$$

where the last expression uses the meaning of f as a linear function $f: U \rightarrow \mathbb{C}$. Note that $\{ \psi, \tilde{\psi} \}$ agrees with the anticommutator of the field operators, $\psi \tilde{\psi} + \tilde{\psi} \psi$.

Now recall that the quantum dynamics is determined by a Hamiltonian H that acts on ψ by the commutator $H \cdot \psi = [H, \psi]$. The one-parameter groups $t \mapsto e^{itH/\hbar}$ generated by this action (the time evolutions) preserve the symmetric pairing:

$$\{ \psi, \tilde{\psi} \} = \{ e^{itH/\hbar} \psi, e^{itH/\hbar} \tilde{\psi} \}$$

since the anticommutation relations [1] do not change with time. They also preserve the unitary structure,

$$\langle \psi, \tilde{\psi} \rangle = \langle e^{itH/\hbar} \psi, e^{itH/\hbar} \tilde{\psi} \rangle$$

because probability in Nambu space is conserved. (Physically speaking, this holds true as long as H is quadratic, i.e., many-body interactions are negligible.)

One can now pose Dyson’s question again: given Nambu space V and a symmetry group G acting on it, what is the set of time evolution operators that preserve the structure of V and are compatible with G ? From the section “The framework,” we know the answer to be some symmetric space, but which are the symmetric spaces that occur?

Class D

Consider a superconductor with no symmetries in its quasiparticle dynamics, so $G = \{\text{Id}\}$. (A concrete example would be a disordered spin-triplet superconductor in the vortex phase.) The time evolutions $U_t = e^{itH/\hbar}$ are then constrained only by invariance of the unitary structure and the symmetric pairing $\{, \}$ of Nambu space. These two structures are consistent; they are related by particle–hole conjugation C :

$$\{ \psi, \tilde{\psi} \} = \langle C\psi, \tilde{\psi} \rangle$$

which is an antiunitary operator with square $C^2 = +\text{Id}$. Let $V_R \subset V$ denote the real vector space of fixed points of C . (The field operators in V_R are called of “Majorana” type in physics.) The condition $\{ \psi, \tilde{\psi} \} = \{ U_t \psi, U_t \tilde{\psi} \}$ selects a complex orthogonal

group $\text{SO}(V)$, and imposing unitarity yields a real orthogonal subgroup $\text{SO}(V_R)$ with $\dim V_R \in 2\mathbb{N}$ – a symmetric space of the D family.

When expressed in some basis of Majorana fermions (meaning a basis of V_R), the matrix of the time evolution generator $iH \in \mathfrak{so}(V_R)$ is real skew, and that of H imaginary skew. The simplest random matrix model for class D , the SO -invariant Gaussian ensemble of imaginary skew matrices, is analyzed in the second edition of Mehta’s (1991) book. From the expressions given by Mehta it is seen that the level correlation functions at high energy coincide with those of the Wigner–Dyson universality class of unitary symmetry. The level correlations at low energy, however, show different behavior defining a separate universality class. This universal behavior at low energies has immediate physical relevance, as it is precisely the low-energy quasiparticles that determine the thermal transport properties of the superconductor at low temperatures.

Class DIII

Let now magnetic fields and magnetic impurities be absent, so that time reversal T is a symmetry of the quasiparticle system: $G = \{\text{Id}, T\}$. Following the section “The framework,” the set of good time evolutions is $M \cong K/K_\tau$ with $K = \text{SO}(V_R)$ and K_τ the set of fixed points of $U \mapsto \tau(U) = TUT^{-1}$ in K . What is K_τ ?

The square of the time-reversal operator is $T^2 = -\text{Id}$ (for particles with spin $1/2$), and commutes with particle–hole conjugation C , which makes $P := iCT$ a useful operator to consider. Since C by definition commutes with the action of K , and hence also with that of K_τ , the subgroup K_τ has an equivalent description as

$$K_\tau = \{ k \in U(V) \mid k = PkP^{-1} = \tau(k) \}$$

The operator P is easily seen to have the following properties: (1) P is unitary, (2) $P^2 = \text{Id}$, and (3) $\text{tr}_V P = 0$. Consequently, P possesses two eigenspaces V_\pm of equal dimension, and the condition $k = PkP^{-1}$ fixes a subgroup $U(V_+) \times U(V_-)$ of $U(V)$. Since P contains a factor $i = \sqrt{-1}$ in its definition, it anticommutes with the antilinear operator T . Therefore, the automorphism τ exchanges $U(V_+)$ with $U(V_-)$, and the fixed-point set K_τ is the same as $U(V_+) \cong U_{2N}$. Thus,

$$M = K/K_\tau \cong \text{SO}_{4N}/U_{2N} \quad (\dim V_+ = 2N)$$

a symmetric space in the $DIII$ family. Note that for particles with spin $1/2$ the dimension of V_+ has to be even.

By realizing the algebra of involutions C, T as $C\psi = (1_{2N} \otimes i\sigma_x)\bar{\psi}$ and $T\psi = (1_{2N} \otimes i\sigma_y)\bar{\psi}$, the Hamiltonians H in class DIII are brought into the standard form

$$H = \begin{pmatrix} 0 & Z \\ -\bar{Z} & 0 \end{pmatrix}$$

where the $2N \times 2N$ matrix Z is complex and skew.

Class C

Next let the spin of the quasiparticles be conserved, as is the case for a spin-singlet superconductor with no spin-orbit scatterers present, and let time-reversal invariance be broken by a magnetic field. The symmetry group of the quasiparticle system then is the spin-rotation group: $G = G_0 = \text{Spin}_3 = \text{SU}_2$.

Nambu space V can be arranged to be a tensor product $V = L \otimes R$ so that G_0 acts trivially on L and by the spinor representation on the spinor space $R \cong \mathbb{C}^2$. Since two spinors combine to give a scalar, the latter comes with an alternating bilinear form $a: R \times R \rightarrow \mathbb{C}$. In a suitable basis, the anticommutation relations [1] factor on particle-hole and spin indices. The symmetric bilinear form $\{, \}$ of V correspondingly factors under the tensor product decomposition $V = L \otimes R$ as

$$\{l_1 \otimes r_1, l_2 \otimes r_2\} = [l_1, l_2] \times a(r_1, r_2)$$

where $[,]$ is an alternating form on L , giving L the structure of a complex symplectic vector space.

The good set M now consists of the time evolutions that, in addition to preserving the structure of Nambu space, commute with the spin-rotation group SU_2 :

$$M = \{U \in \text{U}(V) | UC = CU, \forall g \in \text{SU}_2 : gU = Ug\}$$

By the last condition, all time evolutions act trivially on the factor R . The condition $UC = CU$, which expresses invariance of the symmetric form of V , then implies that time evolutions preserve the alternating form of L . Time evolutions therefore are unitary symplectic transformations of L , hence $M = \text{USp}(L) \cong \text{USp}_{2N}$ – a symmetric space of the C family. The Hamiltonian matrices in class C have the standard form

$$H = \begin{pmatrix} W & Z \\ Z^\dagger & -\bar{W} \end{pmatrix}$$

with W being Hermitian and Z complex and symmetric.

Class CI

The next class is obtained by taking the time reversal T as well as the spin rotations $g \in \text{SU}_2$ to be symmetries of the quasiparticle system.

By arguments that should be familiar by now, the set of good time evolutions is a symmetric space $M \cong K/K_\tau$ with $K = \text{USp}(L)$ and K_τ the set of fixed points of τ in K . Once again, the question to be answered is: what is K_τ ? The situation here is very similar to the one for class DIII, with L and $\text{USp}(L)$ taking the roles of V and $\text{SO}(V_R)$. By adapting the previous argument to the present case, one shows that K_τ is the same as $\text{U}(L_+) \cong \text{U}_N$, where L_+ is the positive eigenspace of $P = iCT$ viewed as a unitary operator on L . Thus,

$$M = K/K_\tau \cong \text{USp}_{2N}/\text{U}_N$$

Dirac Fermions: The Chiral Classes

Three large families of symmetric spaces remain to be implemented. Although these, too, occur in mesoscopic physics, their most natural realization is by 4D Dirac fermions in a random gauge field background.

Consider the Lagrangian \mathcal{L} for the Euclidean spacetime version of QCD with $N_c \geq 3$ colors of quarks coupled to an SU_{N_c} gauge field A_μ :

$$\mathcal{L} = i\bar{\psi} \gamma^\mu (\partial_\mu - A_\mu) \psi + im\bar{\psi} \psi$$

The massless Dirac operator $D = i\gamma^\mu (\partial_\mu - A_\mu)$ anti-commutes with $\gamma_5 = \gamma^0 \gamma^1 \gamma^2 \gamma^3$. Therefore, in a basis of eigenstates of γ_5 the matrix of D takes the form

$$D = \begin{pmatrix} 0 & Z \\ Z^\dagger & 0 \end{pmatrix} \quad [3]$$

If the gauge field carries topological charge $\nu \in \mathbb{Z}$, the Dirac operator D has at least $|\nu|$ zero modes by the index theorem. To make a simple model of the challenging situation where A_μ is distributed according to Yang-Mills measure, one takes the matrices Z to be complex rectangular, of size $p \times q$ with $p - q = \nu$, and puts a Gaussian probability measure on that space. This random matrix model for D captures the universal features of the QCD Dirac spectrum in the massless limit.

The exponential of the truncated Dirac operator, e^{itD} (where t is not the time), lies in a space equivalent to $\text{U}_{p+q}/\text{U}_p \times \text{U}_q$ – a symmetric space of the AIII family. We therefore say that the universal behavior of the QCD Dirac spectrum is that of symmetry class AIII.

But hold on! Why are we entitled to speak of a symmetry class here? By definition, symmetries

always *commute* with the Hamiltonian, never do they anticommute! (The relation $D = -\gamma_5 D \gamma_5$ is not a symmetry in the sense of Dyson, nor is it a symmetry in our sense.)

Class AIII

To incorporate the massless QCD Dirac operator into the present classification scheme, we adapt it to the Nambu space setting. This is done by reorganizing the four-component Dirac spinor $\psi, \bar{\psi}$ as an eight-component Majorana spinor Ψ , to write

$$\mathcal{L}_{m=0} = \frac{i}{2} \Psi \Gamma^\mu (\partial_\mu - \mathcal{A}_\mu) \Psi$$

The 8×8 matrices Γ^μ are real symmetric besides satisfying the Clifford relations $\Gamma^\mu \Gamma^\nu + \Gamma^\nu \Gamma^\mu = 2\delta^{\mu\nu}$. A possible tensor product realization is

$$\begin{aligned} \Gamma^0 &= 1 \otimes \sigma_z \otimes 1, & \Gamma^1 &= \sigma_x \otimes \sigma_y \otimes \sigma_y \\ \Gamma^2 &= \sigma_y \otimes \sigma_y \otimes 1, & \Gamma^3 &= \sigma_z \otimes \sigma_y \otimes \sigma_y \end{aligned}$$

The gauge field in this Majorana representation is $\mathcal{A}_\mu = 1 \otimes 1 \otimes (A_\mu^{(-)} - A_\mu^{(+)}) \sigma_y$ where $A_\mu^{(\pm)} = (1/2)(A_\mu \pm A_\mu^\dagger)$ are the symmetric and skew parts of $A_\mu \in \mathfrak{su}(N_c)$.

The operator $H = i\Gamma^\mu (\partial_\mu - \mathcal{A}_\mu)$ is imaginary skew, therefore e^{iH} is real orthogonal. This means that there exists a Nambu space V with unitary structure \langle, \rangle and symmetric pairing $\{, \}$, both of which are preserved by the action of e^{iH} . No change of physical meaning or interpretation is implied by the identical rewriting from Dirac D to Majorana H . The fact that Dirac fermions are not truly Majorana is encoded in a U_1 -symmetry $H e^{i\theta Q} = e^{i\theta Q} H$ generated by $Q = 1 \otimes 1 \otimes \sigma_y$.

Now comes the essential point: since H obeys $\bar{H} = -H$, the chiral “symmetry” $H = -\Gamma_5 H \Gamma_5$ with $\Gamma_5 = 1 \otimes \sigma_x \otimes 1$ can be recast as a true symmetry:

$$H = +\Gamma_5 \bar{H} \Gamma_5 = T H T^{-1}$$

with antilinear $T: \Psi \mapsto \Gamma_5 \bar{\Psi}$. Thus, the massless QCD Dirac operator is indeed associated with a symmetry class in the present, post-Dyson sense: that is class AIII, realized by self-adjoint operators

on Nambu space with Dirac U_1 -symmetry and an antiunitary symmetry T .

Classes BDI and CII

Consider Hamiltonians D still of the form [3] but now with matrix entries taken from either the real numbers or the real quaternions. Their one-parameter groups e^{iD} belong to two further families of symmetric spaces, namely the classes BDI and CII of Table 1. These large families are known to be realized as symmetry classes by the massless Dirac operator with gauge group SU_2 (for BDI), or with fermions in the adjoint representation (for CII). For the details we must refer to Verbaarschot’s (1994) paper and the recent article by Heinzner *et al.* (2005).

See also: Classical Groups and Homogeneous Spaces; Compact Groups and Their Representations; Determinantal Random Fields; Dirac Fields in Gravitation and Nonabelian Gauge Theory; Dirac Operator and Dirac Field; High T_c Superconductor Theory; Integrable Systems in Random Matrix Theory; Lie Groups: General Theory; Random Matrix Theory in Physics; Random Partitions; Supersymmetry Methods in Random Matrix Theory; Symmetries and Conservation Laws.

Further Reading

Altland A, Simons BD, and Zirnbauer MR (2002) Theories of low-energy quasi-particle states in disordered d -wave superconductors. *Physics Reports* 359: 283–354.
Altland A and Zirnbauer MR (1997) Non-standard symmetry classes in mesoscopic normal-/superconducting hybrid systems. *Physical Review B* 55: 1142–1161.
Dyson FJ (1962) The threefold way: algebraic structure of symmetry groups and ensembles in quantum mechanics. *Journal of Mathematical Physics* 3: 1199–1215.
Heinzner P, Huckleberry A, and Zirnbauer MR (2005) Symmetry classes of disordered fermions. *Communications in Mathematical Physics* 257: 725–771.
Helgason S (1978) *Differential Geometry, Lie Groups and Symmetric Spaces*. New York: Academic Press.
Mehta ML (1991) *Random Matrices*. New York: Academic Press.
Weyl H (1939) *The Classical Groups: Their Invariants and Representations*. Princeton: Princeton University Press.
Verbaarschot JJM (1994) The spectrum of the QCD Dirac operator and chiral random matrix theory: the threefold way. *Physical Review Letters* 72: 2531–2533.

Synchronization of Chaos

M A Aziz-Alaoui, Université du Havre, Le Havre, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction: Chaotic Systems Can Synchronize

Synchronization is a ubiquitous phenomenon characteristic of many processes in natural systems and (nonlinear) science. It has permanently remained an objective of intensive research and is today considered as one of the basic nonlinear phenomena studied in mathematics, physics, engineering, or life science. This word has a Greek root, *syn* = common and *chronos* = time, which means to share the common time or to occur at the same time, that is, correlation or agreement in time of different processes (Boccaletti *et al.* 2002). Thus, synchronization of two dynamical systems generally means that one system somehow traces the motion of another. Indeed, it is well known that many coupled oscillators have the ability to adjust some common relation that they have between them due to weak interaction, which yields to a situation in which a synchronization-like phenomenon takes place.

The original work on synchronization involved periodic oscillators. Indeed, observations of (periodic) synchronization phenomena in physics go back at least as far as C Huygens (1673), who, during his experiments on the development of improved pendulum clocks, discovered that two very weakly coupled pendulum clocks become synchronized in phase: two clocks hanging from a common support (on the same beam of his room) were found to oscillate with exactly the same frequency and opposite phase due to the (weak) coupling in terms of the almost imperceptible oscillations of the beam generated by the clocks.

Since this discovery, periodic synchronization has found numerous applications in various domains, for instance, in biological systems and living nature where synchronization is encountered on different levels. Examples range from the modeling of the heart to the investigation of the circadian rhythm, phase locking of respiration with a mechanical ventilator, synchronization of oscillations of human insulin secretion and glucose infusion, neuronal information processing within a brain area and communication between different brain areas. Also, synchronization plays an important role in several neurological diseases such as epilepsies and pathological tremors, or in different forms of cooperative

behavior of insects, animals, or humans (Pikovsky *et al.* 2001).

This process may also be encountered in celestial mechanics, where it explains the locking of revolution period of planets and satellites.

Its view was strongly broadened with the developments in radio engineering and acoustics, due to the work of Eccles and Vincent, 1920, who found synchronization of a triode generator. Appleton, Van der Pol, and Van der Mark, 1922–27, have, experimentally and theoretically, extended it and worked on radio tube oscillators, where they observed entrainment when driving such oscillators sinusoidally, that is, the frequency of a generator can be synchronized by a weak external signal of a slightly different frequency.

But, even though original notion and theory of synchronization implies periodicity of oscillators, during the last decades, the notion of synchronization has been generalized to the case of interacting chaotic oscillators. Indeed, the discovery of deterministic chaos introduced new types of oscillating systems, namely the chaotic generators.

Chaotic oscillators are found in many dynamical systems of various origins; the behavior of such systems is characterized by instability and, as a result, limited predictability in time.

Roughly speaking, a system is chaotic if it is deterministic, has a long-term aperiodic behavior, and exhibits sensitive dependence on initial conditions on a closed invariant set (the chaos theory is discussed in more detail elsewhere in this encyclopedia) (*see* Chaos and Attractors).

Consequently, for a chaotic system, trajectories starting arbitrarily close to each other diverge exponentially with time, and quickly become uncorrelated. It follows that two identical chaotic systems cannot synchronize. This means that they cannot produce identical chaotic signals, unless they are initialized at exactly the same point, which is in general physically impossible. Thus, at first sight, synchronization of chaotic systems seems to be rather surprising because one may intuitively (and naively) expect that the sensitive dependence on initial conditions would lead to an immediate breakdown of any synchronization of coupled chaotic systems. This scenario in fact led to the belief that chaos is uncontrollable and thus unusable. Despite this, in the last decades, the search for synchronization has moved to chaotic systems. Significant research has been done and, as a result, Yamada and Fujisaka (1983), Afraimovich *et al.* (1986), and Pecora and Carroll (1990) showed that

two chaotic systems could be synchronized by coupling them: synchronization of chaos is actual and chaos could then be exploitable. Ever since, many researchers have discussed the theory and the design or applications of synchronized motion in coupled chaotic systems. A broad variety of applications has emerged, for example, to increase the power of lasers, to synchronize the output of electronic circuits, to control oscillations in chemical reactions, or to encode electronic messages for secure communications.

The publication of the seminal paper of Pecora and Carroll (1990) had a very strong impact in the domain of chaos theory and chaos synchronization, and their applications. It had stimulated very intense research activities and the related studies continue to attract great attention. Many authors have contributed to developing this domain, theoretically or experimentally (Boccaletti *et al.* 2002, Pecorra *et al.* 1997, references therein).

However, the special features of chaotic systems make it impossible to directly apply the methods developed for synchronization of periodic oscillators. Moreover, in the topics of coupled chaotic systems, many different phenomena, which are usually referred to as synchronization, exist and have been studied now for over a decade. Thus, more precise descriptions of such systems are indeed desirable.

Several different regimes of synchronization have been investigated. In the following, the focus will be on explaining the essentials on this large topic, subdivided into four basic types of synchronization of coupled or forced chaotic systems which have been found and have received much attention, while emphasizing on the first three:

- identical (or complete) synchronization (IS), which is defined as the coincidence of states of interacting systems;
- generalized synchronization (GS), which extends the IS phenomenon and implies the presence of some functional relation between two coupled systems; if this relationship is the identity, we recover the IS;
- phase synchronization (PS), which means entrainment of phases of chaotic oscillators, whereas their amplitudes remain uncorrelated; and
- lag synchronization (LS), which appears as a coincidence of time-shifted states of two systems.

Other regimes exist, some of them will be briefly pointed out at the end of this article; we also will briefly discuss the very relevant issue of the stability of synchronous motions.

Our discussion and examples given here are based on unidirectionally continuous systems, most of the exposed ideas can be easily extended to discrete systems.

Let us also emphasize that the same year, 1990, saw the publication of another seminal paper, by Ott, Grebogi, and Yorke (OGY) on the control of chaos (Ott *et al.* 1990). Recently, it has been realized that synchronization and control of chaos share a common root in nonlinear control theory. Both topics were presented by many authors in a unified framework. However, synchronization of chaos has evolved in its own right, even if it is nowadays known as a part of the nonlinear control theory.

Synchronization and Stability

For the basic master–slave configuration, where an autonomous chaotic system (the master)

$$\frac{dX}{dt} = F(X), \quad X \in \mathbb{R}^n \quad [1]$$

drives another system (the slave),

$$\frac{dY}{dt} = G(X, Y), \quad Y \in \mathbb{R}^m \quad [2]$$

synchronization takes place when Y asymptotically copies, in a certain manner, a subset X_p of X . That is, there exists a relation between the two coupled systems, which could be a smooth invertible function ψ , which transforms the trajectories on the attractor of a first system into those on the attractor of a second system. In other words, if we know, after a transient regime, the state of the first system, it allows us to predict the state of the second: $Y(t) = \psi(X(t))$. Generally, it is assumed that $n \geq m$; however, for the sake of easy readability (even if this is not a necessary restriction) the case $n = m$ will only be considered; thus, $X_p = X$. Henceforth, if we denote the difference $Y - \psi(X)$ by X_\perp , in order to arrive at a synchronized motion, it is expected that

$$\|X_\perp\| \rightarrow 0, \quad \text{as } t \rightarrow +\infty \quad [3]$$

If ψ is the identity function, the process is called IS.

Definition of IS System [2] synchronizes with system [1], if the set $M = \{(X, Y) \in \mathbb{R}^n \times \mathbb{R}^n, Y = X\}$ is an attracting set with a basin of attraction $B(M \subset B)$ such that $\lim_{t \rightarrow \infty} \|X(t) - Y(t)\| = 0$, for all $(X(0), Y(0)) \in B$.

Thus, this regime corresponds to the situation where all the variables of two (or more) coupled chaotic systems converge.

If ψ is not the identity function, the phenomenon is more general and is referred to as GS.

Definition of GS System [2] synchronizes with system [1], in the generalized sense, if there exists a transformation $\psi: \mathbb{R}^n \rightarrow \mathbb{R}^m$, a manifold $M = \{(X, Y) \in \mathbb{R}^{n+m}, Y = \psi(X)\}$ and a subset B ($M \subset B$), such that for all $(X_0, Y_0) \in B$, the trajectory based on the initial conditions (X_0, Y_0) approaches M as time goes to infinity. This is explained further in the following.

Henceforth, in the case of IS, eqn [3] above means that a certain hyperplane M , called synchronization manifold, within \mathbb{R}^{2n} , is asymptotically stable. Consequently, for the sake of synchrony motion, we have to prove that the origin of the transverse system $X_\perp = Y - X$ is asymptotically stable. That is, to prove that the motion transversal to the synchronization manifold dies out.

However, significant progress has been made by mathematicians and physicists in studying the stability of synchronous motions. Two main tools are used in the literature for this aim: conditional Lyapunov exponents and asymptotic stability. In the examples given below, we will essentially formulate conditions for synchronization in terms of Lyapunov exponents, which play a central role in chaos theory. These quantities measure the sensitive dependence on initial conditions for a dynamical system and also quantify synchronization of chaos.

The Lyapunov exponents associated with the variational equation corresponding to the transverse system X_\perp :

$$\frac{dX_\perp}{dt} = DF(X)X_\perp \quad [4]$$

where $DF(X)$ is the Jacobian of the vector field evaluated onto the driving trajectory X , are referred to as transverse or conditional Lyapunov exponents (CLEs).

In the case of IS, it appears that the condition $L_{\max}^\perp < 0$ is sufficient to insure synchronization, where L_{\max}^\perp is the largest CLE. Indeed, eqn [4] gives the dynamics of the motion transverse to the synchronization manifold; therefore, CLEs indicate if this motion dies out or not, and hence, whether the synchronization state is stable or not. Consequently, if L_{\max}^\perp is negative, it insures the stability of the synchronized state. This will be best explained using two examples below.

Even if there exist other approaches for studying synchronization, one may ask if this condition on L_{\max}^\perp is true in general. To answer this question, mathematicians have recently formulated it in terms of properties of manifolds (or synchronization hyperplanes). Some rigorous results on (generalized)

synchronization, when the system is smooth, are given by Josic (2000). This approach relies on the Fenichel theory of normally hyperbolic invariant manifolds and quantities that resemble Lyapunov exponents, and is referred to as differentiable GS. However, many situations correspond to the case where, in some region of values of parameters coupling, the function ψ is only continuous but not smooth, that is, the graph of ψ is a complicated geometrical object. This kind of synchronization is called nonsmooth GS (Afraimovich *et al.* 2001).

Furthermore, the mathematical theory of IS often assumes the coupled oscillators to be identical, even if, in practice, no two oscillators are exact copies of each other. This leads to small differences in system parameters and then to synchronization errors. These errors have been studied by many authors (see, e.g., Illing (2002), and references therein).

Identical Synchronization

Perhaps the best way to explain synchronization of chaos is through IS, also referred to as conventional or complete synchronization (Boccaletti *et al.* 2002). It is the simplest form of chaos synchronization and generalizes to the complete replacement which is explained below. It is also the most typical form of chaotic synchronization often observable in two identical systems.

There are various processes leading to synchronization; depending on the particular coupling configuration used these processes could be very different. So, one has to distinguish between the following two main situations, even if they are, in some sense, similar: the unidirectional and the bidirectional coupling. Indeed, synchronization of chaotic systems is often studied for schemes of the form

$$\begin{aligned} \frac{dX}{dt} &= F(X) + kN(X - Y) \\ \frac{dY}{dt} &= G(Y) + kM(X - Y) \end{aligned} \quad [5]$$

where F and G act in \mathbb{R}^n , $(X, Y) \in (\mathbb{R}^n)^2$, is a scalar, and M and N are coupling matrices belonging to $\mathbb{R}^{n \times n}$. If $F = G$ the two subsystems X and Y are identical. Moreover, when both matrices are non-zero then the coupling is called bidirectional, while it is referred to as unidirectional if one is the zero matrix, and the other nonzero.

Constructing Pairs of Synchronized Systems: Complete Replacement

Pecora and Carroll (1990) proposed the use of stable subsystems of given chaotic systems to

construct pairs of unidirectionally coupled synchronizing systems. Since then generalizations of this approach have been developed and various methods now exist to synchronize systems (Wu 2002, Hasler 1998).

One way to build a couple of synchronized systems is then to use the basic construction method introduced by Pecora and Carroll, who made an important observation. They found that, when they make a replica of part of a chaotic system and send a system variable from the original system (transmitter) to drive this replica (receiver), sometimes the replica subsystem and the original chaotic one lock in their steps and evolve together chaotically in synchrony. This method can be described as follows. Consider the autonomous n -dimensional dynamical system,

$$\frac{du}{dt} = F(u) \tag{6}$$

divide this system into two subsystems ($u = (v, w)$),

$$\begin{aligned} \frac{dv}{dt} &= G(v, w) \\ \frac{dw}{dt} &= H(v, w) \end{aligned} \tag{7}$$

where $v = (u_1, \dots, u_m)$, $w = (u_{m+1}, \dots, u_n)$, $G = (F_1, \dots, F_m)$, and $H = (F_{m+1}, \dots, F_n)$. Next, create a new subsystem w' identical to the w -subsystem. This yields a $(2n - m)$ -dimensional system:

$$\begin{aligned} \frac{dv}{dt} &= G(v, w) \\ \frac{dw}{dt} &= H(v, w) \\ \frac{dw'}{dt} &= H(v, w') \end{aligned} \tag{8}$$

The first state-variable component $v(t)$ of the (v, w) system is then used as the input to the w' -system. The coupling is unidirectional and the (v, w) subsystem is referred to as the driving (or master) system, the w' -subsystem as the response (or slave) system. In this context, the following notions and results are useful.

Definition If $\lim_{t \rightarrow +\infty} \|w'(t) - w(t)\| = 0$ and $w'(t)$ continues to remain in step with $w(t)$ in the course of the time, the two subsystems are said to be synchronized.

Definition The Lyapunov exponents of the response subsystem (w') for a particular driven trajectory $v(t)$ are called CLEs.

Let $w(t)$ be a chaotic trajectory with initial condition $w(0)$, and $w'(t)$ be a trajectory started at a nearby point $w'(0)$. The basic idea of the Pecora–Carroll approach is to establish the asymptotic stability of the solutions of w' -subsystem by means of CLEs. They have shown the following result (Pecora and Carroll 1990):

Theorem A necessary and sufficient condition for the two subsystems, w and w' , to be synchronized is that all of the CLEs be negative.

Note that only a finite number of possible decompositions (or couplings) v – w exist; this is bounded by the number of different possible subsystems, namely $N(N - 1)/2$. (For a description and mathematical analysis of various coupling schemes see Wu (2002).) Furthermore, by splitting the main system [6] in a different way, (complete) synchronization could not exist. Indeed, in general, only a few of the possible response subsystems possess negative CLEs, and may thus be used to implement synchronizing systems using the Pecora–Carroll method. In fact, it has been pointed out in the literature that in some cases, the CLE criterion is not as practical as some other criteria.

For simplicity, the idea will now be developed on the following three-dimensional simple autonomous system, which belongs to the class of dynamical systems called generalized Lorenz systems (see Derivière and Aziz-Alaoui (2003), and references therein):

$$\begin{aligned} \dot{x} &= -9x - 9y \\ \dot{y} &= -17x - y - xz \\ \dot{z} &= -z + xy \end{aligned} \tag{9}$$

(This should be compared with the well-known Lorenz system:

$$\begin{aligned} \dot{x} &= -10x + 10y \\ \dot{y} &= 28x - y - xz \\ \dot{z} &= -\frac{8}{3}z + xy \end{aligned}$$

which differs in the signs of various terms and the values of coefficients.) From previous observations, it was shown that system [9] oscillates chaotically; its Lyapunov exponents are +0.601, 0.000, and –16.470; it exhibits the chaotic attractor of Figure 1, with a three-dimensional feature very similar to that of Lorenz attractor (in fact, it satisfies the condition $z < 0$, but in our context it does not matter).

Let us divide system [9] into two subsystems $v = x_1$ and $w = (y_1, z_1)$. By creating a copy

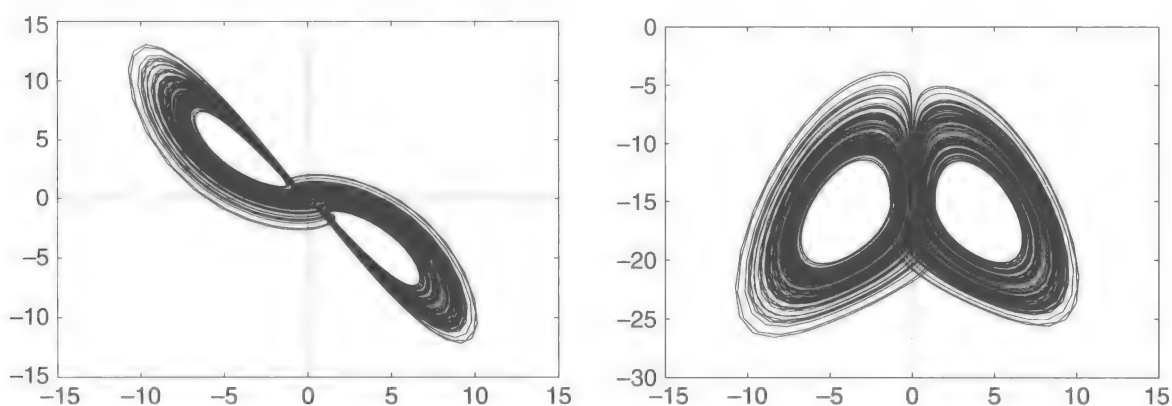


Figure 1 The chaotic attractor of system [9]: x - y and x - z plane projections.

$w'=(y_2, z_2)$ of the w -subsystem, we obtain the following five-dimensional dynamical system:

$$\begin{aligned} \dot{x}_1 &= -9x_1 - 9y_1 \\ \dot{y}_1 &= -17x_1 - y_1 - x_1z_1 \\ \dot{z}_1 &= -z_1 + x_1y_1 \\ \dot{y}_2 &= -17x_1 - y_2 - x_1z_2 \\ \dot{z}_2 &= -z_2 + x_1y_2 \end{aligned} \tag{10}$$

In numerical experiments, it was observed that the motion quickly results in the two equalities, $\lim_{t \rightarrow +\infty} |y_2 - y_1| = 0$ and $\lim_{t \rightarrow +\infty} |z_2 - z_1| = 0$, to be satisfied, that is, $\lim_{t \rightarrow +\infty} \|w' - w\| = 0$. These equalities persist as the system evolves. Hence, the two subsystems w and w' are synchronized. Figure 2 illustrates this phenomenon.

It is also easy to verify that the synchronization persists even if a slight change in the parameters of the system is made. The CLEs of the linearization of the system around the synchronous state, the negativity of which determines the stability of the synchronized solution, are also computed easily.

Pecora–Carroll similarly built the system [10] by using the following steps. Starting with two copies of system [9], a signal $x(t)$ is transmitted from the first to the second: in the second system all x -components are replaced with the signal from the first system, that is, x_2 is replaced by x_1 in the second system. Finally, the dx_2/dt equation is eliminated, since it is exactly the same as dx_1/dt equation, and is superfluous. This then results in system [10]. For this reason, Pecora–Carroll called this construction a complete replacement. Thus, it is natural to think of the x_1 variable as driving the second system, but also to label the first system the drive and the second system the response. In fact, this method is a particular case of the unidirectional coupling method explained below. Note also that this method could be modified by using a partial substitution approach, in which a response variable

is replaced with the drive counterpart only in certain locations (Pecora *et al.* 1997).

Unidirectional IS

The IS synchronization has also been called as one-way diffusive coupling, drive–response coupling, master–slave coupling, or negative feedback control.

System [5], $F = G$ and $N = 0$, becomes unidirectionally coupled, and reads

$$\begin{aligned} \frac{dX}{dt} &= F(X) \\ \frac{dY}{dt} &= F(Y) + kM(X - Y) \end{aligned} \tag{11}$$

M is then a matrix that determines the linear combination of X components that will be used in the difference, and k determines the strength of the coupling (see, for an interesting review on this subject, Pecora *et al.* (1997)). In unidirectional synchronization, the evolution of the first system (the drive) is unaltered by the coupling, the second system (the response) is then constrained to copy the dynamics of the first. Let us consider an example with two copies of system [9], and for

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{12}$$

that is, by adding a damping term to the first equation of the response system, we get a following unidirectionally coupled system, coupled through a linear term $k > 0$ according to variables $x_{1,2}$:

$$\begin{aligned} \dot{x}_1 &= -9x_1 - 9y_1 \\ \dot{y}_1 &= -17x_1 - y_1 - x_1z_1 \\ \dot{z}_1 &= -z_1 + x_1y_1 \\ \dot{x}_2 &= -9x_2 - 9y_2 - k(x_2 - x_1) \\ \dot{y}_2 &= -17x_2 - y_2 - x_2z_2 \\ \dot{z}_2 &= -z_2 + x_2y_2 \end{aligned} \tag{13}$$

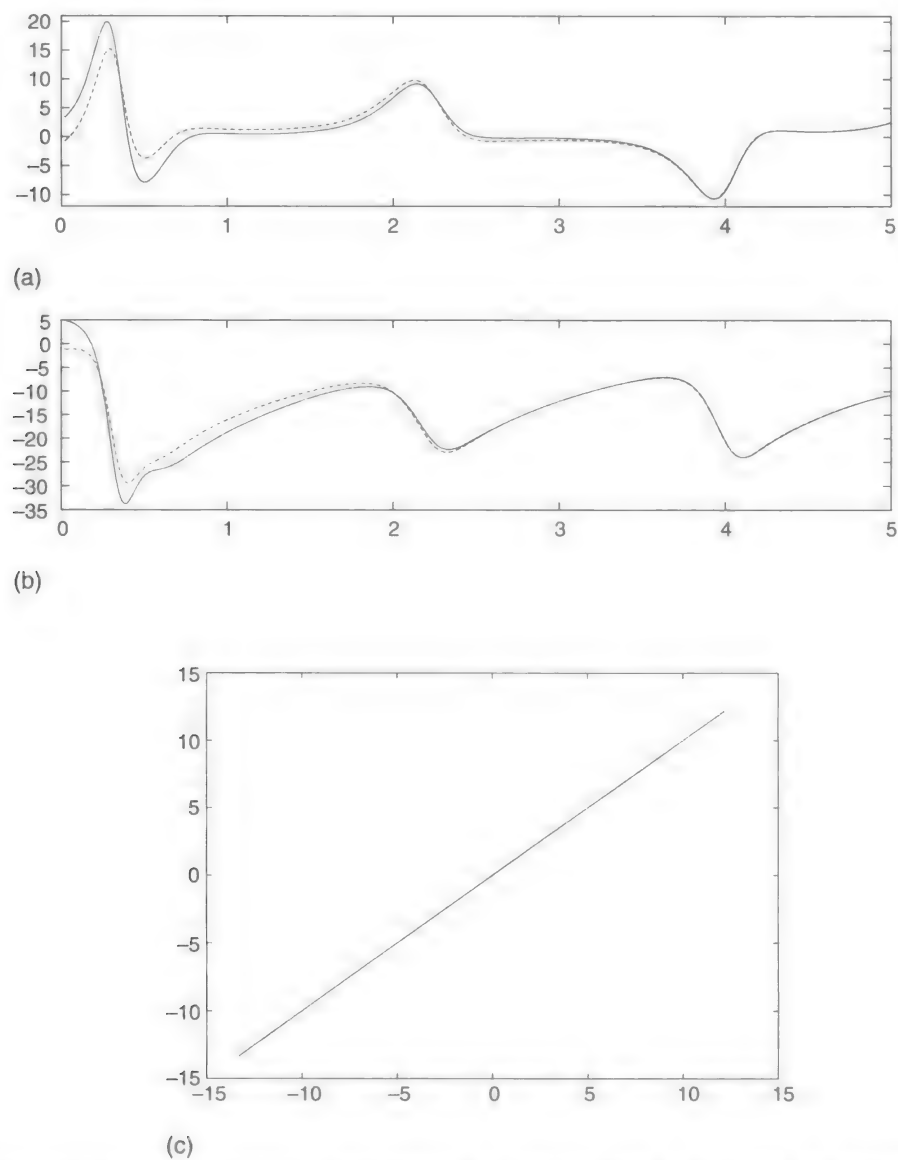


Figure 2 Complete replacement synchronization. Time series for (a) $y_i(t)$ and (b) $z_i(t)$, $i = 1, 2$, in system [10]. The difference between the variable of the transmitter and the variable of the receiver asymptotes tends to zero as time progresses, that is, synchronization occurs after transients die down. (c) The plot of amplitudes y_1 against y_2 , after transients die down, shows a diagonal line, which also indicates that the receiver and the transmitter are maintaining synchronization. The plot of z_1 against z_2 shows a similar behavior.

For $k = 0$, the two subsystems are uncoupled; for $k > 0$ both subsystems are unidirectionally coupled; and for $k \rightarrow +\infty$, we recover the complete replacement coupling scheme explained above. Our numerical computations yield the optimal value \tilde{k} for the synchronization; we found that for $k \geq \tilde{k} = 4.999$, both subsystems of [13] synchronize. That is, starting from random initial conditions, and after some transient time, system [13] generates the same attractor as for system [9] (see Figure 1). Consequently, all the variables of the coupled chaotic subsystems converge: x_2 converges to x_1 , y_2 to y_1 , and z_2 to z_1 (see Figure 3). Thus, the second system (the response) is locked to the first one (the drive).

Alternatively, observation of diagonal lines in correlation diagrams, which plot the amplitudes x_1

against x_2 , y_1 against y_2 , and z_1 against z_2 , can also indicate the occurrence of system synchronization.

IS was the first for which examples of unidirectionally coupled chaotic systems were presented. It is important for potential applications of chaos synchronization in communication systems, or for time-series analysis, where the information flow is also unidirectional.

Bidirectional IS

A second brief example uses a bidirectional (also called mutual or two-way) coupling. In this situation, in contrast to the unidirectional coupling, both drive and response systems are connected in such a way that they influence each other's behavior. Many

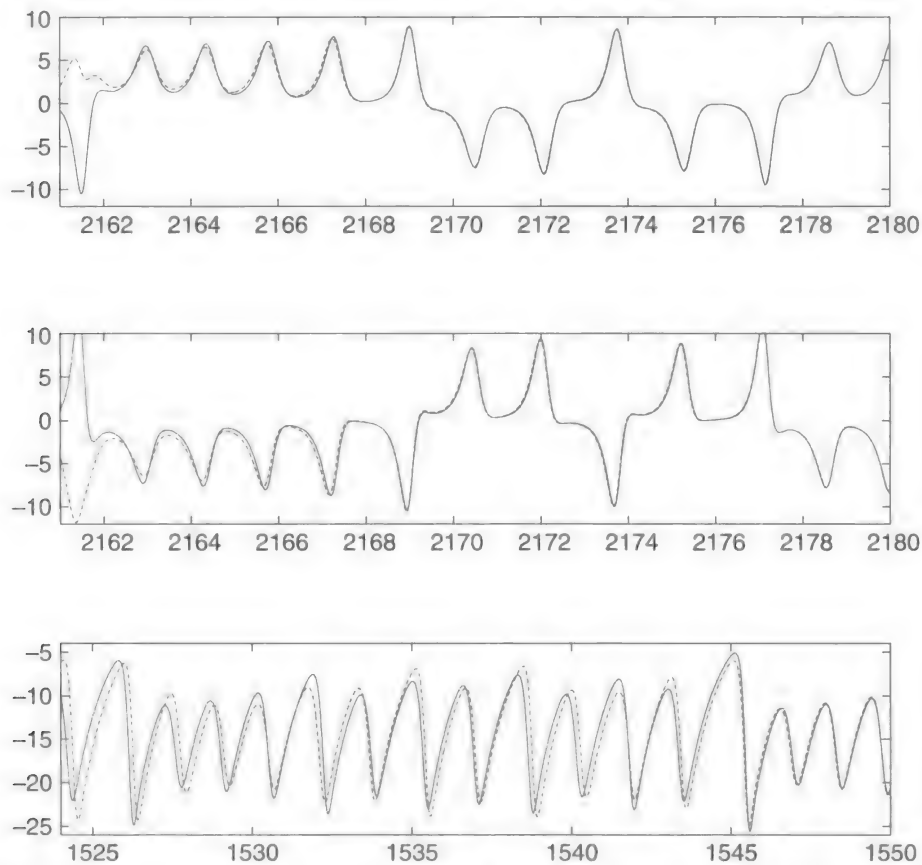


Figure 3 Time series for $x_i(t)$, $y_i(t)$, and $z_i(t)$ ($i = 1, 2$) in system [13] for the coupling constant $k = 5.0$, that is, beyond the threshold necessary for synchronization. After transients die down, the two subsystems synchronize perfectly.

biological or physical systems consist of bidirectionally interacting elements or components; examples range from cardiac and respiratory systems to coupled lasers with feedback. Let us then take two copies of the same system [9] as given above, but two-way coupled through a linear constant term $k > 0$ according to variables $x_{1,2}$:

$$\begin{aligned}
 \dot{x}_1 &= -9x_1 - 9y_1 - k(x_1 - x_2) \\
 \dot{y}_1 &= -17x_1 - y_1 - x_1z_1 \\
 \dot{z}_1 &= -z_1 + x_1y_1 \\
 \dot{x}_2 &= -9x_2 - 9y_2 - k(x_2 - x_1) \\
 \dot{y}_2 &= -17x_2 - y_2 - x_2z_2 \\
 \dot{z}_2 &= -z_2 + x_2y_2
 \end{aligned}
 \tag{14}$$

We can get an idea of the onset of synchronization by plotting, for example, x_1 against x_2 for various values of the coupling-strength parameter k . Our numerical computations yield the optimal value \tilde{k} for the synchronization: $\tilde{k} \simeq 2.50$ (Figure 4), both (x_i, y_i, z_i) subsystems synchronize and system [14] also generates the attractor of Figure 1.

Synchronization manifold and stability Geometri-Geometrically, the fact that systems [13] and [14], beyond synchronization, generate the same attractor

as system [9], implies that the attractors of these combined drive-response six-dimensional systems are confined to a three-dimensional hyperplane (the synchronization manifold) defined by $Y = X$. After the synchronization is reached, this manifold is a stable submanifold in the full phase space \mathbb{R}^6 . Figure 5 gives an idea of what the geometry of the synchronous attractor of system [13] or [14] looks like, by exhibiting the projection of the phase space \mathbb{R}^6 onto (x_1, y_1, y_2) subspace. But, one can similarly plot any combination of variable x_i , y_i , and z_i ($i = 1, 2$), and get the same result, since the motion, in case of synchronization, is confined to the hyperplane defined in \mathbb{R}^6 by the equalities $x_1 = x_2$, $y_1 = y_2$, and $z_1 = z_2$.

This hyperplane is stable since small perturbations which take the trajectory off the synchronization manifold decay in time. Indeed, as stated earlier, CLEs of the linearization of the system around the synchronous state could determine the stability of the synchronized solution. This leads to requiring that the origin of the transverse system, X_\perp , is asymptotically stable. To see this, for both systems [13] and [14], we then switch to the new set of coordinates, $X_\perp = Y - X$, that is, $x_\perp = x_2 - x_1$, $y_\perp = y_2 - y_1$, and $z_\perp = z_2 - z_1$. The origin $(0, 0, 0)$ is obviously a fixed point for this transverse system,

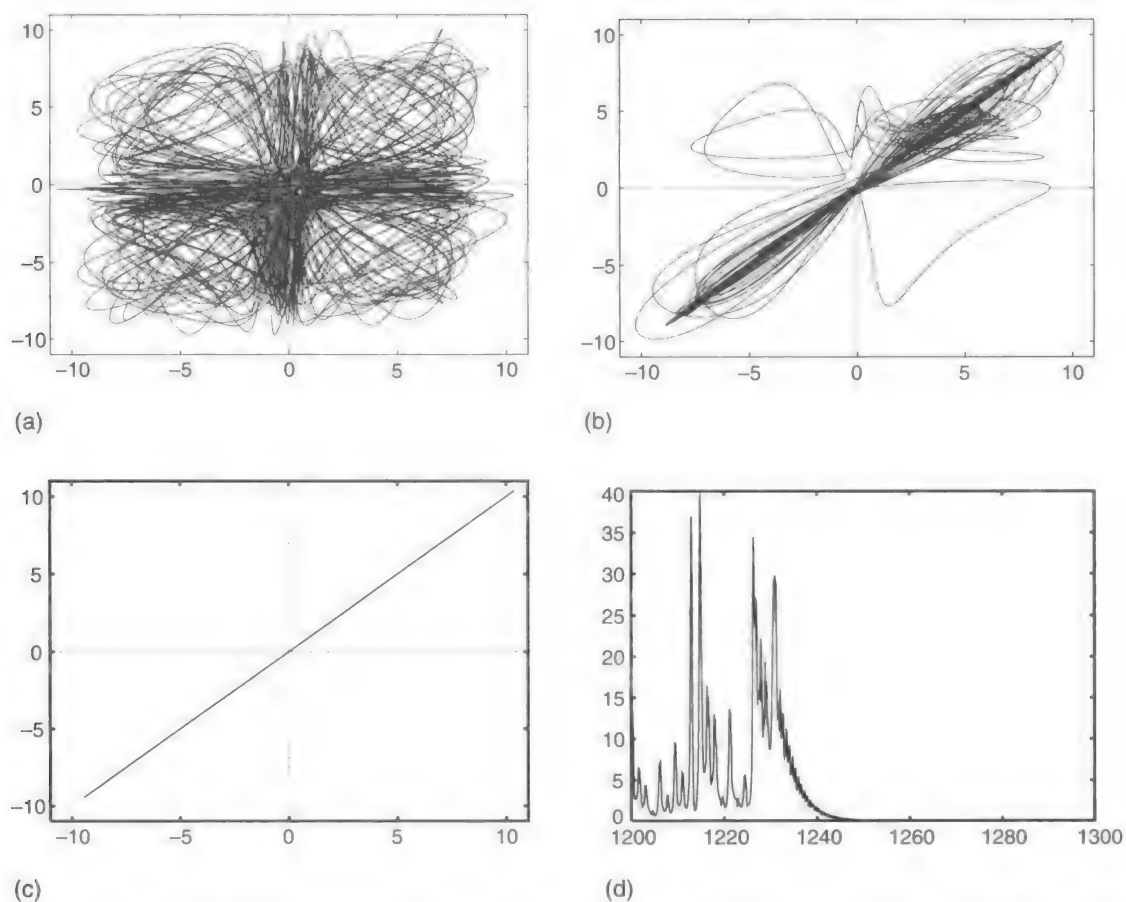


Figure 4 Illustration of the onset of synchronization of system [14]. (a)–(c) Plots of amplitudes x_1 against x_2 for values of the coupling parameter $k = 0.5, 1.5, 2.8$, respectively. The system synchronizes for $k \geq 2.5$. (d) Plot, for $k = 2.8$, of the norm $N(X) = \|x_1 - x_2\| + \|y_1 - y_2\| + \|z_1 - z_2\|$ versus t , which shows that the system synchronizes very quickly.

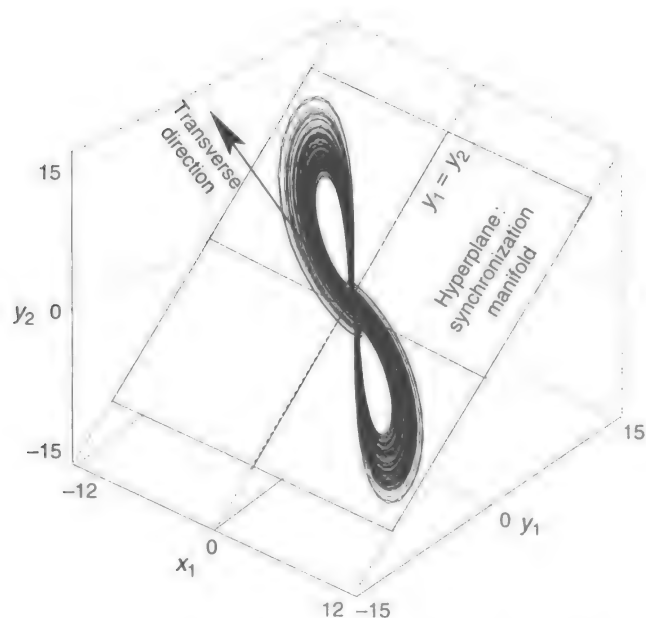


Figure 5 The motion of synchronized system [13] or [14] takes place on a chaotic attractor which is embedded in the synchronization manifold, that is, the hyperplane defined by $x_1 = x_2$, $y_1 = y_2$, and $z_1 = z_2$.

within the synchronization manifold. Therefore, for small deviations from the synchronization manifold, this system reduces to a typical variational equation:

$$\frac{dX_{\perp}}{dt} = DF(X)X_{\perp} \tag{15}$$

where $DF(X)$ is the Jacobian of the vector field evaluated onto the driving trajectory X , that is,

$$\begin{pmatrix} \frac{dx_{\perp}}{dt} \\ \frac{dy_{\perp}}{dt} \\ \frac{dz_{\perp}}{dt} \end{pmatrix} = V \begin{pmatrix} x_{\perp} \\ y_{\perp} \\ z_{\perp} \end{pmatrix} \tag{16}$$

For systems [13] and [14], we obtain

$$V = V_i = \begin{pmatrix} -9 - k_i & -9 & 0 \\ -17 - z & -1 & -x \\ y & x & -1 \end{pmatrix} \tag{17}$$

with $k_i = k$ for system [13] and $k_i = 2k$ for system [14]. Let us remark that the only difference between both matrices V_i is the coupling k which has a factor

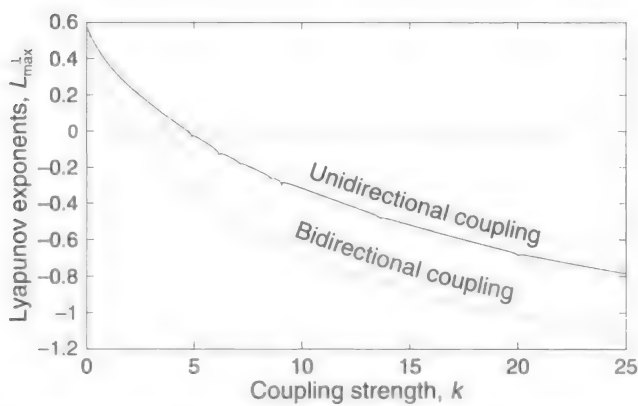


Figure 6 The largest transverse Lyapunov exponents L_{\max}^{\perp} as a function of coupling strength k , in the unidirectional system [13] (solid) and the bidirectional system [14] (dotted).

2 in the bidirectional case. Figure 6 shows the dependence of L_{\max}^{\perp} on k , for both examples of unidirectionally and bidirectionally coupling systems. L_{\max}^{\perp} becomes negative as k increases, which insures the stability of the synchronized state for systems [13] and [14].

Let us note that this can also be proved analytically as done by Derivière and Aziz-Alaoui (2003) by using a suitable Lyapunov function, and using some new extended version of LaSalle invariance principle.

Desynchronization motion Synchronization depends not only on the coupling strength, but also on the vector field and the coupling function. For some choice of these quantities, synchronization may occur only within a finite range $[k_1, k_2]$ of coupling strength; in such a case a desynchronization phenomenon occurs. Thus, increasing k beyond the critical value k_2 yields loss of the synchronized motion (L_{\max}^{\perp} becomes positive).

Generalized Synchronization

Identical chaotic systems synchronize by following the same chaotic trajectory. However, real systems are in general not identical. For instance, when the parameters of two coupled identical systems do not match, or when these coupled systems belong to different classes, complete IS may not be expected, because there does not exist such an invariant manifold $Y = X$, as for IS. For non-identical systems, the possibility of some type of synchronization has been investigated (Afraimovich *et al.* 1986). It was shown that when two different systems are coupled with sufficiently strong coupling strength, a general synchronous relation between their states could exist and it could be

expressed by a smooth invertible function, $Y(t) = \psi(X(t))$. This phenomenon, called GS, is thus a relaxed and extended form of IS in non-identical systems.

However, it may also occur for pairs of identical systems, for example, for systems having reflection symmetry, $F(-X) = -F(X)$. Besides these examples of GS, others also exist that exploit symmetries of the underlying systems (Parlitz and Kocarev 1999).

GS was introduced for unidirectionally coupled systems by Rulkov *et al.* (1995). For simplicity, we also focus on unidirectionally coupled continuous time systems:

$$\begin{aligned} \frac{dX}{dt} &= F(X) \\ \frac{dY}{dt} &= G(Y, u(t)) \end{aligned} \quad [18]$$

where $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$, $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$, $G: \mathbb{R}^m \times \mathbb{R}^k \rightarrow \mathbb{R}^m$, and $u(t) = (u_1(t), \dots, u_k(t))$ with $u_i(t) = h_i(X(t, X_0))$. Two (non-identical) dynamical systems are said to be synchronized in a generalized sense if there is a continuous function ψ from the phase space of the first to the phase space of the second, taking orbits of the first system to orbits of the second.

The main problem is to know when and under what conditions system [18] undergoes GS. Many authors have addressed this question, and it has been shown that asymptotic stability is equally significant for this more universal concept (for some theoretical results, see Rulkov *et al.* (1995) and Parlitz and Kocarev (1999)). For unidirectionally coupled continuous time systems, the following results hold:

Theorem A necessary and sufficient condition for system [18] to be synchronized in the generalized sense is that for each $u(t) = u(X(t, X_0))$ the system- is asymptotically stable.

When it is not possible to find a Lyapunov function in order to use this theorem, one can numerically compute the CLEs of the response system, and use the following result:

Theorem The drive and response subsystems of system [18] synchronize in the generalized sense iff all of the CLEs of the response subsystem are negative.

The definition of ψ has the advantage that it allows the discussion of synchronization of non-identical systems and, at the same time, to consider synchronization in terms of the property of synchronization manifold. Therefore, it is important to study the existence of the transformation ψ and its nature

(continuity, smoothness, ...). Unfortunately, except in special cases (Afraimovich *et al.* 1986), rarely will one be able to produce formulas exhibiting the mapping ψ .

An example of two unidirectionally coupled chaotic systems which synchronize in the generalized sense is given below. Consider the following Rössler system driven by system [9]:

$$\begin{aligned} \dot{x}_1 &= -9x_1 - 9y_1 \\ \dot{y}_1 &= -17x_1 - y_1 - x_1z_1 \\ \dot{z}_1 &= -z_1 + x_1y_1 \\ \dot{x}_2 &= -y_2 - z_2 - k(x_2 - (x_1^2 + y_1^2)) \\ \dot{y}_2 &= x_2 + 0.2y_2 - k(y_2 - (y_1^2 + z_1^2)) \\ \dot{z}_2 &= 0.2 + z_2(x_2 - 9.0) - k(z_2 - (x_1^2 + z_1^2)) \end{aligned} \quad [19]$$

As shown in Figure 7, it appears impossible to tell what the relation is between the transmitter subsystem (x_1, y_1, z_1) in eqn [19] and the two Rössler response subsystems (x_2, y_2, z_2) at $k=1$ and $k=100$.

However, GS occurs for large values of the coupling-strength parameter k . Therefore, for such values we expect that orbits of [19] will lie in the vicinity of a certain synchronization manifold. Indeed, let us define the set

$$S = \{(x_1, y_1, z_1, x_2, y_2, z_2) \in \mathbb{R}^6 : x_2 = x_1^2 + y_1^2, y_2 = y_1^2 + z_1^2, z_2 = x_1^2 + z_1^2\}$$

Since the projections of S onto the coordinates (x_1, y_1, x_2) , (y_1, z_1, y_2) , and (x_1, z_1, z_2) are paraboloids, we can see how the synchronization manifold is approached. This is illustrated in Figure 8, where the (x_1, y_1, x_2) projections of typical trajectories are shown at four different coupling values. (See Josic (2000) for other examples and further developments; see also Pecora *et al.* (1997), where the authors summarize a method in order to get an idea

on the functional relation occurring in case of GS, between two coupled systems.)

Phase Synchronization

For coupled non-identical chaotic systems, other types of synchronizations exist. Recently, a rather weak degree of synchronization, the PS, of chaotic systems has been described (Pikovsky *et al.* 2001). The Greek meaning of the word synchronization, mentioned in the introduction, is closely related to this type of processes. The synchronous motion is actually not visible. Indeed, in PS the phases of chaotic systems with PS are locked, that is, there exists a certain relation between them, whereas the amplitudes vary chaotically and are practically uncorrelated. Thus, it is mostly close to synchronization of periodic oscillators.

Definition PS of two coupled chaotic oscillators occurs if, for arbitrary integers n and m , the phase locking condition between the corresponding phases, $|n\phi_1(t) - m\phi_2(t)| \leq \text{constant}$, holds and the amplitudes of both systems remain uncorrelated.

Let us note that such a phenomenon occurs when a zero Lyapunov exponent of the response system becomes negative, while, as explained above, identical chaotic systems synchronize by following the same chaotic trajectory, when their largest transverse Lyapunov exponent of the synchronized manifold decreases from positive to negative values.

Moreover, following the definition above, this phenomenon is best observed when a well-defined phase variable can be identified in both coupled systems. This can be done for strange attractors that spiral around a "hole," or a particular (fixed) point in a two-dimensional projection of the attractor. The typical example is given by the Rössler system, which, for some range of parameters, exhibits a Möbius-

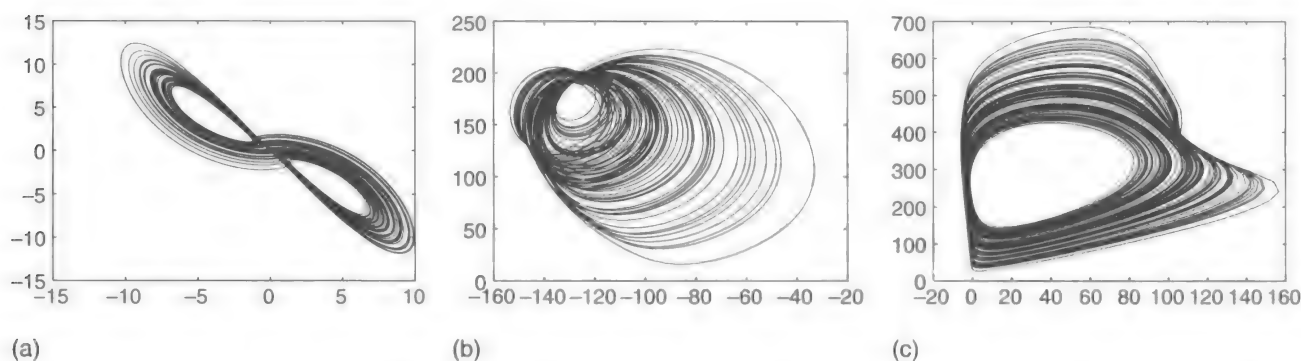


Figure 7 Projections onto the $(x-y)$ plane of typical trajectories of system [19]. (a) (x_1, y_1) projection, that is, a typical trajectory of system [9]; (b) and (c) (x_2, y_2) projections at, respectively, $k=1$ and $k=100$.

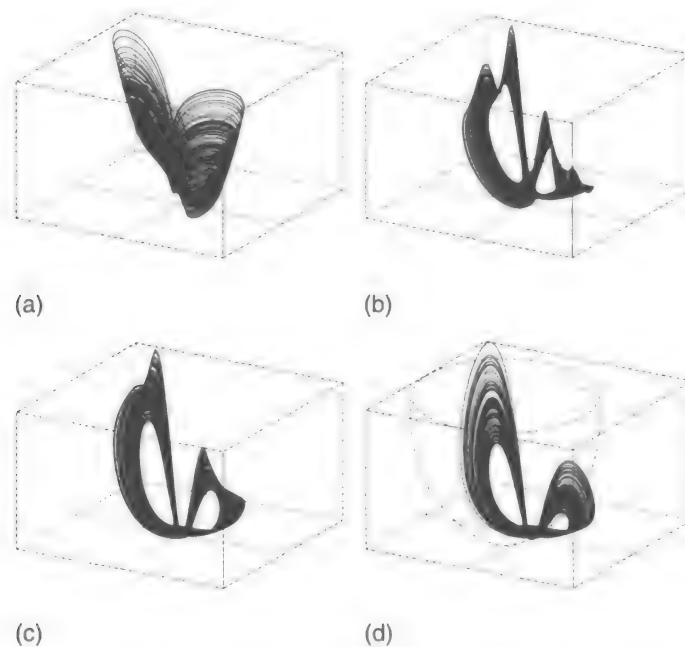


Figure 8 Generalized synchronization. (x_1, y_1, x_2) projections of typical trajectories of system [19] after transients die out, with (a) $k = 1$, (b) $k = 20$, (c) $k = 100$, and (d) $k = 200$. For the last value, the attractor lies in the set S , three-dimensional projections of which are paraboloids.

strip-like chaotic attractor with a central hole. In such a case, a phase angle $\phi(t)$ can be defined that decreases or increases monotonically. For an illustration, we take the following two coupled Rössler oscillators:

$$\begin{aligned}
 \dot{x}_1 &= -\alpha_1 y_1 - z_1 + k(x_2 - x_1) \\
 \dot{y}_1 &= \alpha_1 x_1 + 0.17 y_1 \\
 \dot{z}_1 &= 0.2 + z_1(x_1 - 9.0) \\
 \dot{x}_2 &= -\alpha_2 y_2 - z_2 + k(x_1 - x_2) \\
 \dot{y}_2 &= \alpha_2 x_2 + 0.17 y_2 \\
 \dot{z}_2 &= 0.2 + z_2(x_2 - 9.0)
 \end{aligned} \tag{20}$$

with a small parameter mismatch $\alpha_{1,2} = 0.95 \pm 0.04$, k governs the strength of coupling. If we can define a Poincaré section surface for the system, then, for each piece of a trajectory between two cross sections with this surface, we define the phase, as done in Pikovsky *et al.* (2001), as a piecewise linear function of time, so that the phase increment is 2π at each rotation:

$$\phi(t) = 2\pi \frac{t - t_n}{t_{n+1} - t_n} + 2\pi n, \quad t_n \leq t \leq t_{n+1}$$

where t_n is the time of the n th crossing of the secant surface.

In our example, the last has been chosen as the negative x -axis and represented by the wide segment in Figure 9a. This definition of phases is clearly ambiguous since it depends on the choice of the Poincaré section; nevertheless, defined in this way,

the phase has a physically important property, it does correspond to the direction with the zero Lyapunov exponent in the phase space, its perturbations neither grow nor decay in time. Figure 9c shows that there is a transition from the nonsynchronous phase regime, where the phase difference increases almost linearly with time ($k = 0.01$ and $k = 0.05$), to a synchronous state, where the relation $|\phi_1(t) - \phi_2(t)| < \text{constant}$ holds ($k = 0.1$), that is, the phase difference does not grow with time. However, the amplitudes are obviously uncorrelated as seen in Figure 9b. This example shows that PS could take place for weaker degree of synchronization in chaotic systems. Readers can find more rigorous mathematical discussion on this subject, and on the definition of phases of chaotic oscillators, in Pikovsky *et al.* (2001), see also Boccaletti *et al.* (2002) and references therein.

Other Treatments and Types of Synchronization

Lag Synchronization

PS synchronization occurs when non-identical chaotic oscillators are weakly coupled: the phases are locked, while the amplitudes remain uncorrelated. When the coupling strength becomes larger, some relationships between amplitudes may be established. Indeed, it has been shown (Rosenblum *et al.* 1997), in symmetrically coupled non-identical oscillators and in time-delayed systems, that there exists

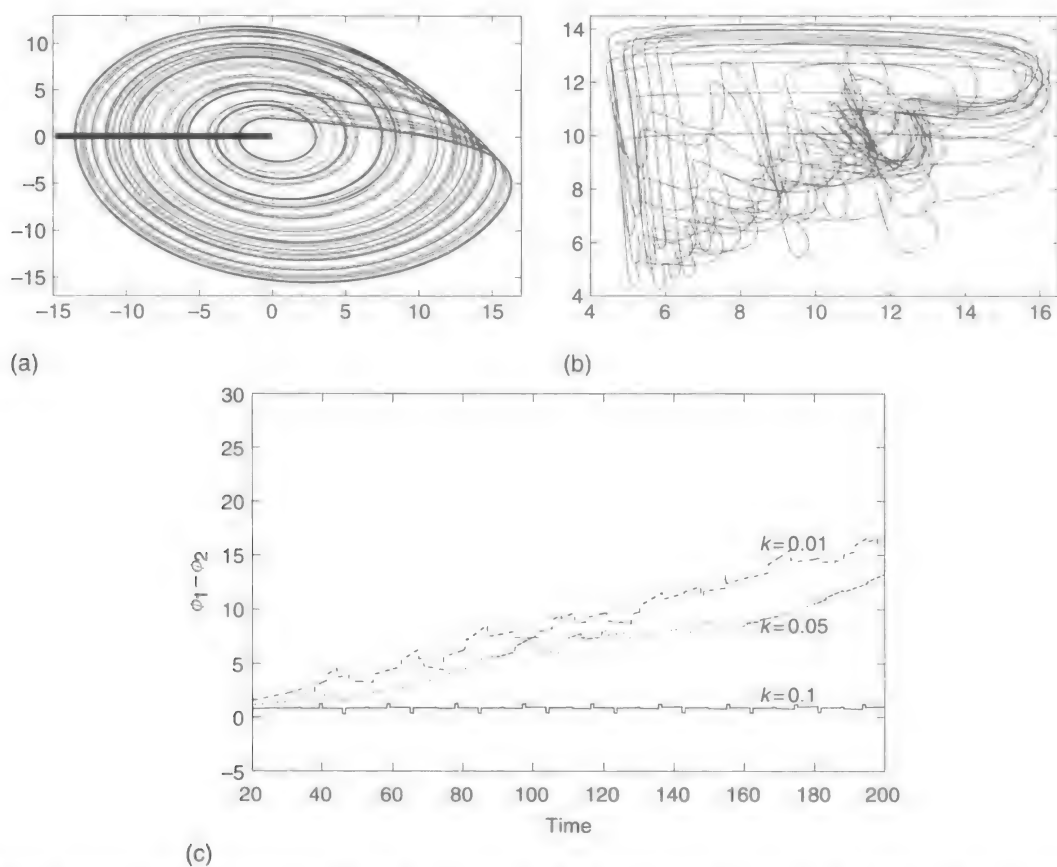


Figure 9 (a) Rössler chaotic attractor projection onto x - y plane. (b) Amplitudes A_1 versus A_2 for the phase synchronized case at $k = 0.1$. (c) Time series of phase difference for different coupling strengths k ; for $k = 0.01$ PS is not achieved, while for $k = 0.1$ PS takes place. Although the phases are locked, for $k = 0.1$, the amplitudes remain chaotic and uncorrelated.

a regime of LS. This process appears as a coincidence of time-shifted states of two systems:

$$\lim_{t \rightarrow +\infty} \|Y(t) - X(t - \tau)\| = 0$$

where τ is a positive delay.

Projective Synchronization

In coupled partially linear systems, it was reported by Mainieri and Rehacek (1999) that two identical systems could be synchronized up to a scaling factor. This type of chaotic synchronization is referred to as projective synchronization. Consider, for example, a three-dimensional chaotic system $\dot{X} = F(X)$, where $X = (x, y, z)$. Decompose X into a vector $v = (x, y)$ and a scalar z ; the system can then be rewritten as

$$\frac{dv}{dt} = g(v, z), \quad \frac{dz}{dt} = h(v, z)$$

In projective synchronization, two identical systems $X_1 = (x_1, y_1, z_1)$ (drive) and $X_2 = (x_2, y_2, z_2)$ (response) are coupled through the scalar variable z . It occurs if the state vectors v_1 and v_2 synchronize up to a constant ratio, that is, $\lim_{t \rightarrow +\infty} \|\alpha v_1(t) - v_2(t)\| = 0$, where α is called a scaling factor. For partially linear systems, it may automatically occur

provided that the systems satisfy some stability conditions.

However, this process could not be classified as GS, even if there exists a linear relation between the coupled systems, because the response system of projective synchronization is not asymptotically stable. For more information about this subject, the reader is referred to Mainieri and Rehacek (1999).

Anticipating Synchronization

It is interesting to mention that a new form of synchronization has recently appeared, the so-called anticipating synchronization (Boccaletti *et al.* 2002). It shows that some coupled chaotic systems might synchronize such that their response anticipates the drivers by synchronizing with their future states.

It is also interesting to mention the nonlinear H_∞ synchronization method for nonautonomous schemes introduced by Suykens *et al.* (1997).

Spatio-Temporal Synchronization

Low-dimensional systems have rather limited usefulness in modeling real-world applications. This is why the synchronization of chaos has been carried

out in high dimensions (see Kocarev *et al.* (1997) for a review). See also Chen and Dong (2001) for a discussion of special high-dimensional systems, namely large arrays of coupled chaotic systems.

Application to Transmission Systems and Secure Communication

Synchronization principles are useful in practical applications. Use of chaotic signals to transmit information has been a very active research topic in the last decade. Thus, it has been established that chaotic circuits may be used to transmit information by synchronization. As a result, several proposals for secure-communication schemes have been advanced (see, e.g., Cuomo *et al.* (1993), Hasler (1998), and Parlitz *et al.* (1999)). The first laboratory demonstration of a secure-communication system, which uses a chaotic signal for masking purposes, and which exploits the chaotic synchronization techniques to recover the signal, was reported by Kocarev *et al.* (1992).

It is difficult, within the scope of this article, to give a complete or detailed discussion, and it should be noted that there exist many competing and tested methods that are well established.

The main idea of the communication schemes is to encode a message by means of a chaotic dynamical system (the transmitter), and to decode it using a second dynamical system (the receiver) that synchronizes with the first. In general, secure-communication applications assume additionally that the coupled systems used are identical.

Different methods can be used to hide the useful information, for example, chaotic masking, chaotic switching, or direct chaotic modulation (Hasler 1998). For instance, in the chaotic masking method, an analog information carrying the signal $s(t)$ is added to the output $y(t)$ of the chaotic system in the transmitter. The receiver tries to synchronize with component $y(t)$ of the transmitted signal $s(t) + y(t)$. If synchronization takes place, the information signal can be retrieved by subtraction (Figure 10).

It is interesting to note that, in all proposed schemes for secure communications using the idea of synchronization (experimental realization or computer simulation), there is an inevitable noise degrading the fidelity of the original message.

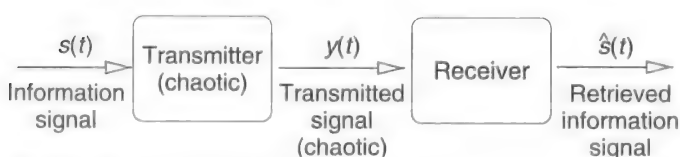


Figure 10 A typical communication setup.

Robustness to parameter mismatch was addressed by many authors (Illing *et al.* 2002). Lozi *et al.* (1993) showed that, by connecting two identical receivers in cascade, a significant amount of the noise can be reduced, thereby allowing the recovery of a much higher quality signal.

Furthermore, different implementations of chaotic secure communication have been proposed during the last decades, as well as methods for cracking this encoding. The methods used to crack such a chaotic encoding make use of the low dimensionality of the chaotic attractors. Indeed, since the properties of low-dimensional chaotic systems with one positive Lyapunov exponent can be reconstructed by analyzing the signal, such as through the delay-time reconstruction methods, it seems unlikely that these systems might provide a secure encryption method. The hidden message can often be retrieved easily by an eavesdropper without using the receiver. But, chaotic masking and encoding are difficult to break, using the state-of-the-art analysis tools, if sufficiently high dimensional chaos generators with multiple positive Lyapunov exponents (i.e., hyperchaotic systems) are used (see Pecora *et al.* (1997), and references therein).

Conclusion

In spite of the essential progress in theoretical and experimental studies, synchronization of chaotic systems continues to be a topic of active investigations and will certainly continue to have a broad impact in the future. Theory of synchronization remains a challenging problem of nonlinear science.

See also: Bifurcations of Periodic Orbits; Chaos and Attractors; Fractal Dimensions in Dynamics; Generic Properties of Dynamical Systems; Isochronous Systems; Lyapunov Exponents and Strange Attractors; Singularity and Bifurcation Theory; Stability Theory and KAM; Weakly Coupled Oscillators.

Further Reading

- Afraimovich V, Chazottes JR, and Cordonet A (2001) Synchronization in directionally coupled systems some rigorous results. *Discrete and Continuous Dynamical Systems B* 1(4): 421–442.
- Afraimovich V, Verichev N, and Rabinovich MI (1986) Stochastic synchronization of oscillations in dissipative systems. *Radio-physics and Quantum Electron* 29: 795–803.
- Boccaletti S, Kurths J, Osipov G, Valladares D, and Zhou C (2002) The synchronization of chaotic systems. *Physics Reports* 366: 1–101.
- Chen G and Dong X (1998) *From Chaos to Order*. Singapore: World Scientific.

- Cuomo K, Oppenheim A, and Strogatz S (1993) Synchronization of Lorenz-based chaotic circuits with applications to communications. *IEEE Transactions on Circuits and Systems – II: Analog and Digital Signal Processing* 40(10): 626–633.
- Derivière S and Aziz-Alaoui MA (2003) Estimation of attractors and synchronization of generalized Lorenz systems. *Dynamics of Continuous, Discrete and Impulsive Systems Series B: Applications and Algorithms* 10(6): 833–852.
- Hasler M (1998) Synchronization of chaotic systems and transmission of information. *International Journal of Bifurcation and Chaos* 8(4): 647–659.
- Huygens Ch (Hugenii) (1673) *Horologium Oscillatorium* (English translation: 1986 *The Pendulum Clock*. Ames: Iowa State University Press). Parisii, France: Apud F. Muguet.
- Illing L (2002) *Chaos Synchronization and Communications in Semiconductor Lasers*. Ph.D. dissertation, San Diego: University of California.
- Illing L, Brocker J, Kocarev L, Parlitz U, and Abarbanel H (2002) When are synchronization errors small? *Physical Review E* 66: 036229.
- Josic K (2000) Synchronization of chaotic systems and invariants manifolds. *Nonlinearity* 13: 1321–1336.
- Kocarev Lj, Halle K, Eckert K, and Chua LO (1992) Experimental demonstration of secure communication via chaotic synchronization. *International Journal of Bifurcation and Chaos* 2(3): 709–713.
- Kocarev Lj, Tasev Z, Stojanovski T, and CParlitz U (1997) Synchronizing spatiotemporel chaos. *Chaos* 7(4): 635–643.
- Lozi R and Chua LO (1993) Secure communications via chaotic synchronization II: noise reduction by cascading two identical receivers. *International Journal of Bifurcation and Chaos* 3(5): 1319–1325.
- Mainieri R and Rehacek J (1999) Projective synchronization in three-dimensional chaotic systems. *Physical Review Letters* 82: 3042–3045.
- Ott E, Grebogi C, and Yorke JA (1990) Controlling chaos. *Physical Review Letters* 64: 1196–1199.
- Parlitz U and Kocarev L (1999) Synchronization of chaotic systems. In: Schuster HG (ed.) *Handbooks of Chaos Control*, pp. 271–303. Germany: Wiley-VCH.
- Pecora L and Carroll T (1990) Synchronization in chaotic systems. *Physical Review Letters* 64: 821–824.
- Pecora L and Carroll T (1991) Driving systems with chaotic signals. *Physical Review A* 44: 2374–2383.
- Pecora L, Carroll T, Johnson G, and Mar D (1997) Fundamentals of synchronization in chaotic systems, concepts and applications. *Chaos* 7(4): 520–543.
- Pikovsky A, Rosenblum M, and Kurths J (2001) *Synchronization, A Universal Concept in Nonlinear Science*. Cambridge: Cambridge University Press.
- Rosenblum M, Pikovsky A, and Kurths J (1997) From phase to lag synchronization in coupled chaotic oscillators. *Physical Review Letters* 78: 4193–4196.
- Rulkov N, Sushchik M, Tsimring L, and Abarbanel H (1995) Generalized synchronization of chaos in directionally coupled chaotic systems. *Physical Review E* 51(2): 980–994.
- Suykens J, Vandewalle J, and Chua LO (1997) Nonlinear H_∞ synchronization of chaotic Lure systems. *International Journal of Bifurcation and Chaos* 7(6): 1323–1335.
- Wu CW (2002) Synchronization in Coupled Chaotic Circuits and Systems. *Series on Nonlinear Science, Series A*, vol. 41. Singapore: World Scientific.
- Yamada T and Fujisaka H (1983) Stability theory of synchronized motion in coupled-oscillator systems. *Progress of Theoretical Physics* 70: 1240–1248.

Thermal Quantum Field Theory

C D Jäkel, Ludwig-Maximilians-Universität München, München, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Quantum field theory was initially invented in order to describe high-energy elementary particles, thereby unifying quantum mechanics and special relativity. In other words, quantum field theory was addressed to the so-called vacuum sector, that is, roughly speaking physics at zero temperature and zero particle density.

The same applies to the various mathematically rigorous versions of quantum field theory that have been developed since the mid-1950s. Indeed, in Wightman's axiomatic setting, quantum field theory is described in terms of a set of the so-called vacuum expectation values. The “algebraic approach” to quantum field theory developed by Araki, Haag, Kastler, and their collaborators is more flexible in nature. In fact, right from the beginning, the new algebraic tools were successfully applied to lattice models and other nonrelativistic systems with infinitely many degrees of freedom (see *Operator algebras and quantum statistical mechanics* by O Bratteli and D W Robinson). But the need to treat large systems of relativistic particles was apparently not felt. Even in Haag's recent monograph, *Local Quantum Physics*, the subjects of algebraic quantum field theory and algebraic quantum statistical mechanics are treated separately.

It is remarkable that constructive field theory was ahead of its time in this respect. The famous $P(\phi)_2$ model (first constructed by Glimm and Jaffe) was adapted to thermal states by Høegh-Krohn as early as 1974 (see Høegh-Krohn (1974)). His paper was properly named “Relativistic quantum statistical mechanics in two-dimensional

space-time,” but only recently has it received proper attention.

At the same time, around 1974, cosmology and heavy-ion collisions drew the interest of physicists towards the quantum statistical mechanics of hot relativistic quantum systems. Well-known papers from this early stage include those by Weinberg, Bernard, and Dolan and Jackiw. While most of the papers used Euclidean path integrals, Umezawa and his school developed a real-time framework called “thermo-field dynamics,” which involved a doubling of the degrees of freedom. The excellent review by Landsman and van Weert (1987) covers these early attempts; it also explains the basic connection to the algebraic approach.

In the following years, it became evident that statistical mechanics (in its standard formulation) is barely sufficient to derive the properties of bulk matter from the underlying microscopic description provided by quantum field theory. Thus, various people began to establish mathematically rigorous foundations for the description of thermal field theory. The most successful approach was launched by D Buchholz (with various collaborators), who, from about 1985 onwards, started applying the KMS condition (which describes a thermal equilibrium state in the operator-algebraic framework of local quantum physics) to relativistic quantum field theory. In 1994, Buchholz and Bros managed to integrate the holomorphic structure of Wightman field theory into Haag's operator-algebraic framework, which led them to the notion of a relativistic KMS condition.

The advanced mathematical concepts involved in the formulation of entropy densities for thermal quantum fields (see Narnhofer (1994)) do not allow us to present this topic. The reader is referred to the excellent book *Quantum Entropy and Its Use* by M Ohya and D Petz for an introduction to the subject. A discussion of the so-called thermalization

effects that occur as a result of a curved spacetime is provided in Quantum Field Theory in Curved Spacetime. Another subject, which is missing almost completely, is perturbation theory. This subject has been covered extensively in three well-known textbooks by Kapusta, Le Bellac, and Umezawa.

Observables and States

Following Heisenberg, we start from the basic assumption that quantum theory can be formulated in terms of observables which form an algebra \mathcal{A} , that is, a vector space with a (noncommutative) multiplication law. Although our emphasis on the abstract algebraic structure may look strange, there is a profound reason for starting out with an abstract algebra of observables: as soon as one considers systems with infinitely many degrees of freedom, one encounters a possibility to realize the abstract elements of the algebra \mathcal{A} as operators on a Hilbert space in various inequivalent ways. The famous equivalence between the Heisenberg and the Schrödinger picture simply breaks down. States which are macroscopically different (e.g., thermal equilibrium states for different temperatures) give rise – in a natural way, which will be discussed in the sequel – to unitarily inequivalent representations of the abstract algebra of observables \mathcal{A} , while states which only differ microscopically can be accommodated by density matrices within the same Hilbert space. In other words, a physical state is described macroscopically by specifying a representation, and microscopically by a density matrix in this representation.

In a Lagrangian approach, the algebra of observables \mathcal{A} may be thought of as being generated by the underlying fields, currents, etc. This leads to the so-called polynomial algebras. It is mathematically convenient to assume that \mathcal{A} is an algebra of bounded operators, generated by the bounded functions of the underlying quantum fields. If $\phi(x)$ is any such field and if $f \in \mathcal{S}(\mathbb{R}^{d+1})$ is any real test function with support in a bounded region of spacetime, then the corresponding operator

$$W(f) = \exp\left(i \int dx f(x) \phi(x)\right)$$

would be a typical element of \mathcal{A} . The set of operators $\{W(f) \mid \text{supp } f \subset \mathcal{O}\}$ will generate a subalgebra $\mathcal{A}(\mathcal{O})$ of \mathcal{A} . The underlying fields can be recovered by taking (functional) derivatives, once a representation of \mathcal{A} on a Hilbert space is specified.

The spacetime symmetry of Minkowski space manifests itself in the existence of a representation

$$\alpha : (\Lambda, x) \mapsto \alpha_{\Lambda, x} \in \text{Aut}(\mathcal{A}), \quad (\Lambda, x) \in \mathcal{P}_+^\dagger$$

of the (orthochronous) Poincaré group \mathcal{P}_+^\dagger . Here $\alpha_{\Lambda, x}$ is an automorphism of \mathcal{A} , that is, a mapping from \mathcal{A} to \mathcal{A} which preserves the algebraic structure. Once a Lorentz frame is fixed by choosing a timelike vector $e \in V_+$, the time evolution $t \mapsto \alpha_{1, te}$ will be denoted by $t \mapsto \tau_t$.

For the free field, the group of automorphisms $(\Lambda, x) \mapsto \alpha_{\Lambda, x}$ is defined by

$$\alpha_{\Lambda, x}(W(f)) := W(f(\Lambda^{-1}(\cdot - x)))$$

As before, $f \in \mathcal{S}(\mathbb{R}^{d+1})$ is a Schwarz function over the Minkowski space \mathbb{R}^{d+1} .

While the invariance of the equations of motion is reflected in the existence of a representation of the Poincaré group in terms of automorphisms in the Heisenberg picture, at least the invariance with respect to Lorentz boosts is spontaneously broken in the Schrödinger picture for a thermal equilibrium state.

The usual notions of vector states and density matrices associated with a given Hilbert space (usually Fock space) are *a priori* not general enough to cover all cases of interest in thermal field theory. The following algebraic definition of a state substantially generalizes the notion of a state: A state ω is a positive, linear, and normalized functional, that is, a linear map $\omega : \mathcal{A} \rightarrow \mathbb{C}$ such that

$$\omega(a^*a) \geq 0 \quad \text{and} \quad \omega(1) = 1$$

Once a state ω is distinguished on physical grounds, the GNS reconstruction theorem provides a Hilbert space \mathcal{H}_ω and a representation π_ω of \mathcal{A} , that is, a map from \mathcal{A} to the set of bounded operators $\mathcal{B}(\mathcal{H}_\omega)$, which preserves the algebraic relations.

It is instructive to consider the GNS representation of the Pauli matrices $\{\sigma_0 = 1, \sigma_1, \sigma_2, \sigma_3\}$. Given a state (a diagonal 2×2 matrix ρ with positive entries and $\text{tr } \rho = 1$), the left regular representation (a construction well known from group theory)

$$\pi(\sigma_i)|\sqrt{\rho}\rangle = |\sigma_i\sqrt{\rho}\rangle, \quad i = 0, 1, 2, 3$$

defines a reducible representation on \mathbb{C}^4 , unless one of the entries in the diagonal of ρ is zero (which corresponds to a pure state). In the latter case, the GNS Hilbert space is \mathbb{C}^2 . By construction, $\langle \sqrt{\rho} | \pi(\sigma_i) | \sqrt{\rho} \rangle = \text{tr } \rho \sigma_i$, $i = 1, 2, 3$.

Thermal Equilibrium

The variety of nonequilibrium states ranges from mild perturbations of equilibrium states through steady states, whose properties are governed by external heat baths, or hydrodynamic flows up to totally chaotic states which no longer

admit a description in terms of thermodynamic notions. Buchholz *et al.* (2002) have initiated an investigation of nonequilibrium states that are locally (but not globally) close to thermal equilibrium. Unfortunately, we will not be able to cover this topic. Instead, we will concentrate on states which deviate from a true equilibrium state only microscopically.

Characterization of Thermal Equilibrium States

When the time evolution $t \mapsto \tau_t \in \text{Aut}(\mathcal{A})$ is changed by a local perturbation, which is slowly switched on and slowly switched off again, then an equilibrium state ω returns to its original form at the end of this procedure. This heuristic condition of adiabatic invariance can be expressed by the stability requirement

$$\lim_{t \rightarrow \infty} \int_{-t}^t dt \, \omega([a, \tau_t(b)]) = 0 \quad \forall a, b \in \mathcal{A} \quad [1]$$

In a pioniering work Haag, Kastler, and Trych-Pohlmeyer showed that the characterization [1] of an equilibrium state leads to a sharp mathematical criterion, first encountered by Haag, Hugenholtz, and Winnink and more implicitly by Kubo, Martin, and Schwinger:

Definition 1 A state ω_β over \mathcal{A} is called a KMS state for some $\beta > 0$, if for all $a, b \in \mathcal{A}$, there exists a function $F_{a,b}$ which is continuous in the strip $0 \leq \Im z \leq \beta$ and analytic and bounded in the open strip $0 < \Im z < \beta$, with boundary values given by

$$F_{a,b}(t) = \omega_\beta(a\tau_t(b)) \quad \text{and} \\ F_{a,b}(t + i\beta) = \omega_\beta(\tau_t(b)a) \quad \forall t \in \mathbb{R} \quad [2]$$

Before we start analyzing the properties of KMS states, we should mention an alternative characterization of thermal equilibrium states: passivity. The amount of work a cycle can perform when applied to a moving thermodynamic equilibrium state is bounded by the amount of work an ideal windmill or turbine could perform; this property is called semipassivity (Kuckert 2002): a state ω is called semipassive (passive) if there is an “efficiency bound” $E \geq 0$ ($E = 0$) such that

$$-(W\Omega_\omega, H_\omega W\Omega_\omega) \leq E \cdot (W\Omega_\omega, |P_\omega| W\Omega_\omega) \\ \forall W \in \pi_\omega(\mathcal{A})''$$

with $W^{-1} = W^*$, $[H_\omega, W] \in \pi_\omega(\mathcal{A})''$, and $[P_\omega, W] \in \pi_\omega(\mathcal{A})''$. Here (H_ω, P_ω) denote the generators implementing the spacetime translations in the GNS representation $(\mathcal{H}_\omega, \Omega_\omega, \pi_\omega)$. Generalizing the notion of complete passivity, the state ω is called completely semipassive if all its finite tensorial powers are

semipassive with respect to one fixed efficiency bound E . It has been shown by Kuckert (2002) that a state is completely semipassive in all inertial frames if and only if it is completely passive in some inertial frame. The latter implies that ω is a KMS state or a ground state (a result due to Puszt and Woronowicz).

Let us now turn to properties of thermal equilibrium states which are specific for relativistic models. It was first recognized by Bros and Buchholz (1994) that KMS states of a relativistic theory have stronger analyticity properties in configuration space than those imposed by the traditional KMS condition:

Definition 2 A KMS state ω_β satisfies the relativistic KMS condition (Bros and Buchholz 1994) if there exists a unit vector e in the forward light cone V_+ such that for every pair of local elements a, b of \mathcal{A} the function $F_{a,b}$

$$F_{a,b}(x_1, x_2) = \omega_\beta(\alpha_{x_1}(a)\alpha_{x_2}(b))$$

extends to an analytic function in the tube domain $-\mathcal{T}_{\beta e/2} \times \mathcal{T}_{\beta e/2}$, where $\mathcal{T}_{\beta e/2} = \{z \in \mathbb{C} \mid \Im z \in V_+ \cap (\beta e/2 - V_+)\}$.

The relativistic KMS condition can be understood as a remnant of the relativistic spectrum condition in the vacuum sector. It has been rigorously established (Bros and Bruchholz 1994) for the KMS states constructed by Buchholz and Junglas (1989) and by C Gérard and the author for the $P(\phi)_2$ model. In the thermal Wightman framework (Bros and Buchholz 1996) it has been shown that the relativistic KMS condition implies existence of model-independent analyticity properties of thermal n -point functions. These properties also appear in perturbative computations of the thermal Wightman functions (Steinmann 1995).

We now turn to the properties of the set of KMS states. For given β , the convex set \mathcal{S}_β of all KMS states is known to form a simplex; the extreme points in the set \mathcal{S}_β are called extremal KMS states. As a consequence, the extremal states in \mathcal{S}_β can be distinguished with the help of “classical” (central) observables, that is, by observables which commute with all other observables.

If ω is an extremal KMS state and γ is an automorphism which commutes with the time evolution $t \mapsto \tau_t$, then the state ω' defined by

$$\omega'(a) := \omega(\gamma(a)), \quad a \in \mathcal{A}$$

is again an extremal KMS state to the same parameter values. If $\omega' \neq \omega$, one says that the symmetry is spontaneously broken.

Lorentz invariance with respect to boosts is always broken by a KMS state, since the KMS condition distinguishes a rest frame. A KMS state might also break spatial translation or rotation invariance. However, by averaging over the different configurations one can usually construct a translation- and rotation-invariant state. The situation is drastically different with respect to supersymmetry. Buchholz and Ojima (1997) have shown that supersymmetry is broken in any thermal state and it is impossible to proceed from it by “symmetrization” to states on which an action of supercharges can be defined.

Existence of Thermal Equilibrium States

Buchholz and Junglas (1989) demonstrated that the existence of KMS states can be guaranteed for a large class of quantum field-theoretic models. The basic assumption to be met concerns the phase-space properties of the model. A generalized trace norm (the so-called “nuclear norm”) is used to estimate the “number” of degrees of freedom in phase space.

The first step is to construct a subspace $\mathcal{H}(\Lambda)$ of the vacuum Hilbert space \mathcal{H}_{vac} , which represents excitations of the vacuum strictly localized inside of a bounded spacetime region $\hat{\mathcal{O}}$. Due to the strong correlations present in the vacuum state of any relativistic model, as a consequence of the Reeh–Schlieder property (see the section “Analyticity of n-point functions”) this is a delicate procedure, which involves the so-called “split property.” This property ensures that there exists a product vector η in vacuum Hilbert space \mathcal{H}_{vac} such that

$$(\eta, \pi_{\text{vac.}}(ab)\eta) = \omega_{\text{vac.}}(a) \cdot \omega_{\text{vac.}}(b) \quad \forall a \in \mathcal{A}(\mathcal{O}), b \in \mathcal{A}(\hat{\mathcal{O}})^c \quad [3]$$

Here $\mathcal{O} \subset \hat{\mathcal{O}}$ denotes a slightly smaller open spacetime region (such that the closure $\bar{\mathcal{O}}$ is inside the interior of $\hat{\mathcal{O}}$) and $\mathcal{A}(\hat{\mathcal{O}})^c := \{A \in \mathcal{A} \mid [A, B] = 0 \ \forall B \in \mathcal{A}(\hat{\mathcal{O}})\}$. The existence of a product vector can be ensured if the nuclear norm satisfies some mild bounds which are expected to hold in all models of physical interest. Given a product vector η which satisfies [3], the sought after subspace is

$$\mathcal{H}(\Lambda) := \overline{\pi_{\text{vac.}}(\mathcal{A}(\mathcal{O}))'' \Omega_{\text{vac.}}}$$

The crucial step in the proof of existence of KMS states is to show that

$$\text{tr } E(\Lambda) e^{-\beta H} E(\Lambda) < \infty \quad \text{for } \beta > 0$$

if the nuclearity condition holds. Here $E(\Lambda)$ denotes the projection onto the subspace $\mathcal{H}(\Lambda)$ representing localized excitations and H denotes the Hamiltonian

in the vacuum representation $\pi_{\text{vac.}}$. Next it is shown that the function

$$\begin{aligned} t &\mapsto \omega_{\beta, \Lambda}(a\tau_t(b)) \\ &= \frac{1}{Z} \text{tr } E(\Lambda) e^{-\beta H} E(\Lambda) \pi_{\text{vac.}}(a\tau_t(b)) \end{aligned}$$

allows an analytic extension to a strip of width β which satisfies the KMS boundary condition [2] for $|t| < \delta$ if $a, b \in \mathcal{A}(\mathcal{O}_\delta)$ and $\mathcal{O}_\delta + te \subset \mathcal{O}$ for $|t| < \delta$. In the final step, Buchholz and Junglas were able to demonstrate that bounds on the nuclear norm are even sufficient to control the thermodynamic limit.

Given a thermal field theory, a slight variation of the method used by Buchholz and Junglas allows one to construct a KMS state for a new temperature (Jäkel 2004), that is, to change the temperature of a thermal state.

Thermal Representations

Given a KMS state ω_β , the GNS construction gives rise to a Hilbert space \mathcal{H}_β and a representation π_β , called a thermal representation, of \mathcal{A} . The algebra $\mathcal{R}_\beta := \pi_\beta(\mathcal{A})''$ possesses a cyclic (due to the GNS construction) and separating (due to the KMS condition) vector Ω_β such that

$$\omega_\beta(a) = (\Omega_\beta, \pi_\beta(a)\Omega_\beta) \quad \forall a \in \mathcal{A}$$

The KMS condition implies that ω_β is invariant under time translations, that is, $\omega_\beta \circ \tau_t = \omega_\beta$ for all $t \in \mathbb{R}$. Thus,

$$U(t)\pi_\beta(a)\Omega_\beta = \pi_\beta(\tau_t(a))\Omega_\beta, \quad a \in \mathcal{A}$$

defines a strongly continuous unitary group $\{U(t)\}_{t \in \mathbb{R}}$ implementing the time evolution in the representation π_β . By Stone’s theorem there exists a self-adjoint generator L such that

$$U(t) = e^{iLt}, \quad t \in \mathbb{R} \quad [4]$$

For $0 \leq \beta < \infty$, the Liouville operator L is not bounded from below; its spectrum is symmetric and consists typically of the whole real line. However, the negative part of L is “suppressed” with respect to the algebra of observables $\mathcal{R}_\beta := \pi_\beta(\mathcal{A})''$ in the following sense (Haag 1992): let $\mathbb{1}_{]-\infty, -\kappa]}$ be the spectral projection of L for the interval $]-\infty, -\kappa] \subset \text{Sp}(L)$, then

$$\|\mathbb{1}_{]-\infty, -\kappa]} A \Omega_\beta\| \leq e^{-\beta\kappa} \|A\| \quad \forall A \in \mathcal{R}_\beta$$

We now turn to structural aspects which are characteristic for a relativistic model, namely the existence of strong spatial correlations and the connection between the decay of these correlations and the spectral properties of the Liouville operator.

Let ω_β be a state, which satisfies the relativistic KMS condition. It follows (using a theorem of Glaser) that for $a \in \mathcal{A}$ the function $\Phi_a: \mathbb{R}^4 \rightarrow \mathcal{H}_\beta$,

$$x \mapsto \pi_\beta(\alpha_x(a))\Omega_\beta$$

can be analytically continued from the real axis into the domain $\mathcal{T}_{\beta e/2}$ such that it is weakly continuous for $\Im z \searrow 0$. If the usual additivity assumption $\cup_i \mathcal{O}_i = \mathcal{O} \Rightarrow \vee_i \mathcal{R}_\beta(\mathcal{O}_i) = \mathcal{R}_\beta(\mathcal{O})$ for the local von Neumann algebras holds, then

$$\mathcal{H}_\beta = \overline{\pi_\beta(\mathcal{A}(\mathcal{O}))\Omega_\beta} \quad [5]$$

for any open spacetime region $\mathcal{O} \subset \mathbb{R}^{d+1}$. Junglas has shown that the thermal Reeh–Schlieder property [5] follows as well from the standard KMS condition, if ω_β is locally normal with respect to the vacuum representation.

The decay of spatial correlations depends on infrared properties of the model, and the essential ingredients for the following cluster theorem are the continuity properties of the spectrum of L near zero.

Theorem 3 *Let Ω_β denote the unique (up to a phase) normalized eigenvector with eigenvalue $\{0\}$ of the Liouvillean L and let P^+ denote the projection onto the strictly positive part of the spectrum of L . Assume that there exist positive constants $m > 0$ and $C_1(\mathcal{O}) > 0$ such that*

$$\begin{aligned} & \|e^{-\lambda L} P^+ \pi_\beta(a) \Omega_\beta\| \\ & \leq C_1(\mathcal{O}) \cdot \lambda^{-m} \|a\| \quad \forall a \in \mathcal{A}(\mathcal{O}) \end{aligned}$$

Here $\mathcal{O} \subset \mathbb{R}^{d+1}$ is an open and bounded spacetime region. Now consider two spacelike separated spacetime regions $\mathcal{O}_1, \mathcal{O}_2$, which can be embedded into \mathcal{O} by translation and such that $\mathcal{O}_1 + \delta e \subset \mathcal{O}'_2, \delta \gg \beta$. then, for $a \in \mathcal{A}(\mathcal{O}_1)$ and $b \in \mathcal{A}(\mathcal{O}_2)$,

$$|\omega_\beta(ba) - \omega_\beta(b)\omega_\beta(a)| \leq C_2 \cdot \delta^{-2m} \|a\| \|b\|$$

The constant $C_2(\beta, \mathcal{O}) \in \mathbb{R}^+$ may depend on the temperature β^{-1} and the size of the region \mathcal{O} but is independent of δ, a , and b .

From explicit calculations one expects that $m = 1/2$ for free massless bosons in $3 + 1$ spacetime dimensions. Consequently, the exponent given on the right-hand side is optimal since it is well known that in this case the correlations decay only like δ^{-1} .

A description of thermal representations would be inadequate without pointing out one of the deepest connections between pure mathematics and physics that emerged in the last century: consider a von Neumann algebra \mathcal{R} which possesses a cyclic and separating vector Ω . Then polar decomposition of the closeable operator $S: A\Omega \mapsto A^*\Omega, A \in \mathcal{R}$, provides an antiunitary operator J (the modular

conjugation) and a self-adjoint operator $\Delta^{1/2}$. The connection to physics was established independently by Takesaki and Winnink, showing that the pair (\mathcal{R}, σ) satisfies the KMS condition for $\beta = -1$, if one sets $\sigma_t(A) = \Delta^{it} A \Delta^{-it}$ for $A \in \mathcal{R}$.

Taking advantage of the Reeh–Schlieder property [5], one can associate modular objects to certain spacetime regions \mathcal{O} . In general, a physical interpretation of these modular objects is missing. But for two-dimensional thermal models, which factorize in light-cone coordinates, the modular group corresponding to the algebra of a spacelike wedge admits a simple description: at large distances (compared to β) from the boundary, the flow pattern is essentially the same as time translations. These are results due to Borchers and Yngvason (1999).

Analyticity Properties of n -Point Functions

The correlation functions describe the full physical content of the theory: all observable quantities can in principle be derived from them. This is so because according to the Wightman reconstruction theorem (which is closely related to the GNS construction) knowledge of the correlation functions allows the reconstruction of the full representation of the field algebra. The Wightman distributions $\{\mathcal{W}_\beta^{(n)}\}_{n \in \mathbb{N}}$,

$$\begin{aligned} & \mathcal{W}_\beta^{(n)}(t_2 - t_1, x_2 - x_1, \dots, t_n - t_{n-1}, x_n - x_{n-1}) \\ & := (\Omega_\beta, \phi_\beta(t_1, x_1) \cdots \phi_\beta(t_n, x_n) \Omega_\beta) \end{aligned} \quad [6]$$

where $\pi_\beta(W(f)) =: \exp(i \int dt dx f(t, x) \phi_\beta(t, x))$, satisfy a number of key properties: locality, positivity, Poincaré covariance, and temperedness. These properties have been formulated for thermal field by Bros and Buchholz (1996), and this section is entirely based on their work.

The relativistic KMS condition implies that the Wightman distributions $\{\mathcal{W}_\beta^{(n)}\}_{n \in \mathbb{N}}$ of a translation-invariant equilibrium state admit in the corresponding set of spacetime variables $(t_2 - t_1, x_2 - x_1), \dots, (t_n - t_{n-1}, x_n - x_{n-1})$ an analytic continuation into the union of domains

$$(\alpha_1 \mathcal{T}_{\beta e}) \times \cdots \times (\alpha_{n-1} \mathcal{T}_{\beta e})$$

for $\alpha_i > 0, i = 1, \dots, n-1$ and $\sum_{i=1}^{n-1} \alpha_i = 1$. The tube domains $\mathcal{T}_{\beta e}$ were specified in Definition 2. For $\beta \rightarrow \infty$, the tube $\mathcal{T}_{\beta e}$ tends to the vacuum tube $\mathcal{T}_{\text{vac.}} = \mathbb{R}^{d+1} + iV_+$; thus, one recovers the spectrum condition for the vacuum expectation values.

Let us now turn to the Fourier transformed Wightman correlation functions. Translation invariance implies

$$\tilde{\mathcal{W}}_\beta^{(n)}(\nu_1, p_1, \dots, \nu_n, p_n) \delta(\nu_1 + \cdots + \nu_n) \delta(p_1 + \cdots + p_n)$$

The Wightman distribution $\tilde{\mathcal{W}}_\beta^{(n)}$ satisfies on the linear manifold $(\nu_1, p_1) + \dots + (\nu_n, p_n) = 0$ the KMS relation in the energy variables: for any pair of multi-indices (I, J) the identity

$$\tilde{\mathcal{W}}_\beta^{(n)}(J, I) = e^{-\beta \nu_I} \tilde{\mathcal{W}}_\beta^{(n)}(I, J)$$

holds, where $\tilde{\mathcal{W}}_\beta^{(n)}(J, I)$ is an abbreviation for $\tilde{\mathcal{W}}_\beta^{(n)}(\{p_i\}_{i \in I}, \{p_j\}_{j \in J})$ and $\nu_I = \sum_{i \in I} \nu_i$.

We now specialize to the two-point function $\mathcal{W}_\beta^{(2)}$. The corresponding commutation function $\mathcal{C}(x)$ is given by

$$\mathcal{C}(x_1 - x_2) = \mathcal{W}_\beta^{(n)}(x_1, x_2) - \mathcal{W}_\beta^{(2)}(x_2, x_1)$$

Locality implies that $\text{supp } \mathcal{C} \subset \overline{V_+} \cup \overline{V_-}$. The retarded and the advanced propagator r and a , formally given by

$$r(x) = i\theta(x_0)\mathcal{C}(x), \quad a(x) = -i\theta(-x_0)\mathcal{C}(x)$$

satisfy the relation

$$r - a = -i\mathcal{C}$$

which corresponds to a partition of the support of \mathcal{C} in its convex components: $\text{supp } r \subset \overline{V_+}$ and $\text{supp } a \subset \overline{V_-}$. For the free scalar field of mass m the commutator function is

$$\mathcal{C}^{(m)}(x) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^4} dp e^{-ipx} \tilde{\mathcal{C}}^{(m)}(p)$$

with

$$\tilde{\mathcal{C}}^{(m)}(p) = \frac{1}{2\pi} \text{sgn}(\nu) \delta(\nu^2 - p^2 - m^2)$$

and subsequently the retarded and advanced propagators $r^{(m)}$ and $a^{(m)}$ are structural functions of the field algebra, which are determined by the c-number commutation relations of the fields. Thus, they are independent of the temperature, in contrast to the two-point function:

$$\tilde{\mathcal{W}}_\beta^{(2)}(p) = \frac{\tilde{\mathcal{C}}^{(m)}(p)}{1 - e^{-\beta \nu}} \quad [7]$$

Let now $\tilde{\tau}(p)$ be the Fourier transform of the time-ordered function $\tau(x)$. The relation

$$\tilde{\tau}(p) = \frac{-i\tilde{r}(p) + -i\tilde{a}(p)e^{-\beta \nu}}{1 - e^{-\beta \nu}}$$

shows that $\tilde{\tau}(p)$ and $-i\tilde{r}(p)$ only “coincide up to an exponential tail” at very high energies (Bros and Buchholz 1996).

Particle Aspects

The condition of locality (together with the relativistic KMS condition) leads to strong constraints on

the general form of the thermal two-point functions that allow one to apply the techniques of the Jost–Lehmann–Dyson representation. As has been shown by Bros and Buchholz (1996), the interacting two-point function \mathcal{W}_β can be represented in the form

$$\mathcal{W}_\beta(t, x) = \int_0^\infty dm \mathcal{D}_\beta(x, m) \mathcal{W}_\beta^{(2)}(t, x, m)$$

Here $\mathcal{D}_\beta(x, m)$ is a distribution in x, m which is symmetric in x , and

$$\mathcal{W}_\beta^{(2)}(t, x, m) = (2\pi)^{-1} \int d\nu dp e^{i(\nu t - px)} \tilde{\mathcal{W}}_\beta^{(2)}(\nu, p)$$

is the two-point correlation function of the free thermal field of mass m . In contrast to the vacuum case, the damping factors $\mathcal{D}_\beta(x, m)$ depend in a nontrivial way on the spatial variables x . The damping factors describe the dissipative effects of the thermal system on the propagation of sharply localized excitations. Bros and Buchholz suggested that the damping factor $\mathcal{D}_\beta(x, m)$ can be decomposed into a discrete and an absolute continuous part

$$\mathcal{D}_\beta(x, m) = \delta(m - m_0) \mathcal{D}_{\beta, d}(x) + \mathcal{D}_{\beta, c}(x, m)$$

and that the δ -contribution in the damping factors is due to stable constituent particles of mass m_0 out of which the thermal states are formed, whereas the collective quasiparticle-like excitations only contribute to the continuous part of the damping factors (Bros and Buchholz 1996).

In the case of spontaneously broken internal symmetries Bros and Buchholz (1998) have shown that the damping factors $\mathcal{D}_\beta^\pm(x, m)$ which appear in the representation of current-field correlations functions

$$\begin{aligned} & (\Omega_\beta, j_0(t, x) \phi_\beta(0, 0) \Omega_\beta) \\ &= \int_0^\infty dm \left(\mathcal{D}_\beta^+(x, m) \partial_t \mathcal{W}_\beta^{(2)}(t, x, m) \right. \\ & \quad \left. + \mathcal{D}_\beta^-(x, m) \mathcal{W}_\beta^{(2)}(t, x, m) \right) \end{aligned}$$

indeed contain a discrete (in the sense of measures) zero-mass contribution and are slowly decreasing in $|x|$ for small values of m . Thus, these damping factors coincide locally with the Källén–Lehmann weights appearing in the case of spontaneous symmetry breaking in the vacuum sector (Bros and Buchholz 1998). It is easily seen in examples that there is no sharp energy–momentum dispersion law for the Goldstone particles. Thus, the Källén–Lehmann representation is better suited than Fourier transformation to uncover the particle aspects of thermal equilibrium states.

Models of Thermal Field Theory

In the simplest case, the classical Lagrangian density of the so-called $P(\phi)_2$ models is given by

$$\mathcal{L} = (\partial_\nu \phi)(\partial^\nu \phi) - m^2 \phi^2 - \frac{\lambda}{4} \phi^4 \quad [8]$$

Here $\phi(t, x)$ denotes a real scalar field over space-time. The construction of the corresponding quantized thermal field presented in this section (Gérard and Jäkel 2005) is based on the original ideas of Høegh-Krohn (1974).

Free Fields

Let \mathfrak{h}_m denote the L^2 -closure of $C_0^\infty(\mathbb{R})$ with respect to the norm $\|f\| = (f, (1/2\epsilon)f)$, where $\epsilon(k) = \sqrt{k^2 + m^2}$ denotes the one-particle energy for a single neutral scalar boson and the scalar product is the usual L^2 -scalar product. The subspaces associated to a double cone \mathcal{O} are given by

$$\mathfrak{h}_m(\mathcal{O}) := \{h \in \mathfrak{h}_m \mid \text{supp } \Re h, \text{supp } \nu^{-1} \Im h \subset \mathcal{O}\}$$

where \mathcal{O} denotes the basis of the double cone \mathcal{O} . The corresponding free quantum field is described by the Weyl algebra $\mathcal{W}(\mathfrak{h}_m) := \{W(f) \mid f \in \mathfrak{h}_m\}$, together with the time evolution $\{\tau_t^\circ\}_{t \in \mathbb{R}}$,

$$\tau_t^\circ(W(f)) = W(e^{it\epsilon}f), \quad f \in \mathfrak{h}_m$$

If $m > 0$, the KMS condition allows just one unique (quasifree) (τ°, β) -KMS state:

$$\omega_\beta^\circ(W(f)) := e^{-(1/4)(f, (1+2\rho)f)_m}, \quad \rho := (e^{\beta\epsilon} - 1)^{-1}$$

The GNS representation associated to the pair $(\mathcal{W}(\mathfrak{h}_m), \omega_\beta^\circ)$ is the well-known Araki–Woods representation, given by

$$\mathcal{H}_{\text{AW}} := \Gamma(\mathfrak{h}_m \oplus \overline{\mathfrak{h}_m}), \quad \Omega_{\text{AW}} := \Omega_{\mathcal{F}},$$

$$\pi_{\text{AW}}(W(h)) := W_{\mathcal{F}}\left((1+\rho)^{1/2}h \oplus \bar{\rho}^{1/2}\bar{h}\right), \quad h \in \mathfrak{h}_m$$

Here $\overline{\mathfrak{h}_m}$ is the Hilbert space conjugate to \mathfrak{h}_m , $W_{\mathcal{F}}(\cdot)$ denotes the usual Weyl operator on the Fock space $\Gamma(\mathfrak{h}_m \oplus \overline{\mathfrak{h}_m})$ and $\Omega_{\mathcal{F}} \in \Gamma(\mathfrak{h}_m \oplus \overline{\mathfrak{h}_m})$ is the Fock vacuum. The Liouvillean L_{AW} (see [4]) can be identified with $d\Gamma(\epsilon \oplus -\epsilon)$.

The local von Neumann algebra generated by $\{\pi_{\text{AW}}(W(h)) \mid h \in \mathfrak{h}_m(\mathcal{O})\}$ is denoted by $\mathcal{R}_{\text{AW}}(\mathcal{O})$. The algebra of observables for the free quantum field (and, as we will see, the $P(\phi)_2$ model) is the norm closure

$$\mathcal{A} := \overline{\bigcup_{\mathcal{O} \subset \mathbb{R}^2} \mathcal{R}_{\text{AW}}(\mathcal{O})}^{\text{C}^*}$$

of the local von Neumann algebras.

The Thermal $P(\phi)_2$ Model

In $1+1$ spacetime dimensions Wick ordering is sufficient to eliminate the UV divergences of polynomial interactions. As it turns out, the leading order in the UV divergences is independent of the temperature (in agreement with the results found in Kopper *et al.* (2001)). Thus, it is a matter of convenience whether one uses the thermal covariance function C_β ,

$$C_\beta(h_1, h_2) := \left(h_1, \frac{(1 + e^{-\beta\epsilon})}{2\epsilon(1 - e^{-\beta\epsilon})} h_2\right)_{L^2(\mathbb{R})}$$

$$h_1, h_2 \in \mathcal{S}(\mathbb{R})$$

or the vacuum covariance function C_{vac} to define the Wick ordering:

$$:\phi(f)^n:_C = \sum_{m=0}^{[n/2]} \frac{n!}{m!(n-2m)!} \phi(f)^{n-2m} \left(-\frac{1}{2}C(f, f)\right)^m$$

Now let $P(\lambda)$ be a real-valued polynomial, which is bounded from below. Then Euclidean techniques can be used to define the operator sum

$$H_l := L_{\text{AW}} + \int_{-l}^l :P(\phi_\beta(x)):_C dx$$

in the Araki–Woods representation and to show that H_l is essentially self-adjoint Gérard and Jäkel (2005). Thus, (the closure of) H_l can be used to define a perturbed time evolution $t \mapsto \tau_t^l$ on \mathcal{A} and the vector

$$\Omega_l := \frac{e^{-(\beta/2)H_l} \Omega_{\text{AW}}}{\|e^{-(\beta/2)H_l} \Omega_{\text{AW}}\|}$$

induces a KMS state ω_l for the dynamical system $(\pi_{\text{AW}}(\mathcal{A})'', \tau^l)$.

A finite propagation speed argument (using Trotter's product formula) shows that

$$\tau_t^l(A) := e^{itH_l} A e^{-itH_l}, \quad t \in \mathbb{R} \quad [9]$$

is independent of l for $A \in \mathcal{R}_{\text{AW}}(\mathcal{O})$, $t \in \mathbb{R}$ fixed and l sufficiently large. Thus, there exists a limiting dynamics τ such that

$$\lim_{l \rightarrow \infty} \|\tau_t^l(A) - \tau_t(A)\| = 0 \quad [10]$$

for all $A \in \mathcal{R}_{\text{AW}}(\mathcal{O})$, \mathcal{O} bounded. This norm convergence extends to the norm closure \mathcal{A} of the local von Neumann algebras.

The existence of weak* limit points (which are states) of the (generalized) sequence $\{\omega_l\}_{l>0}$ is a consequence of the Banach–Alaoglu theorem. The fact that all limit states satisfy the KMS condition with respect to the pair (\mathcal{A}, τ) follows from [10]. To

prove that the sequence $\{\omega_l\}_{l>0}$ has only one accumulation point,

$$\omega_\beta = \lim_{l \rightarrow \infty} \omega_l \quad [11]$$

is more delicate. Following Høegh-Krohn, Nelson symmetry is used in Gérard and Jäkel (2005) to relate the interacting thermal theory on the real line to the $P(\phi)_2$ model on the circle S^1 of length at temperature 0. The existence of the limit [11] then follows from the uniqueness of the vacuum state on the circle. The relativistic KMS condition can be derived by Nelson symmetry as well, using the fact that the discrete spectrum of the model on the circle satisfies the spectrum condition. Since the limit [11] exists on the norm closure \mathcal{A} of the weakly closed local algebras, it follows from a result of Takesaki and Winnink that ω_β is locally normal with respect to the Araki–Woods representation (which itself is locally normal with respect to the Fock representation). Consequently,

$$\mathcal{R}_\beta(\mathcal{O}) := \pi_\beta(\mathcal{A}(\mathcal{O}))'' \cong \mathcal{R}_{\text{AW}}(\mathcal{O}), \quad \mathcal{O} \text{ bounded}$$

that is, $\mathcal{R}_\beta(\mathcal{O})$ is (isomorphic to) the unique hyperfinite factor of type III₁. Moreover, the local Fock property implies that the split property holds.

Perturbation Theory

Steinmann (1995) has shown that perturbative expansions for the Wightman distributions of the $:\phi^4:_4$ model can be derived directly in the thermodynamic limit, using as only inputs the equations of motion and the (thermal) Wightman axioms. The result can be represented as a sum over generalized Feynman graphs.

The method consists in solving the differential equations for the correlation functions which follow from the field equation, by a power series expansion in the coupling constant, using the axiomatic properties of the Wightman functions as subsidiary conditions. The Wightman axioms are expected to hold separately in each order of perturbation theory, with the exception of the cluster property.

As expected, the UV renormalization can be chosen to be temperature independent, that is, one can use the same counterterms as in the vacuum case. But the infrared divergencies are more severe, they cannot be removed by minor adjustments of the renormalization procedure. Various elaborate resummation techniques have been proposed to (at least partially) remove the infrared singularities.

Another approach has been pursued by Koppper *et al.* (2001). They have investigated the perturbation expansion of the $:\phi^4:_4$ model in the imaginary-time formalism, using Wilson's flow equations. The result is once again that all correlation functions become ultraviolet-

finite in all orders of the perturbation expansion, once the theory has been renormalized at zero temperature by usual renormalization prescriptions.

Asymptotic Dynamics of Thermal Fields

Timelike asymptotic properties of thermal correlation functions cannot be interpreted in terms of free fields due to persistent dissipative effects of a thermal system. This well-known fact manifests itself in a softened pole structure of the Green's functions in momentum space and is at the root of the failure of the conventional approach to thermal perturbation theory (Bros and Bruchholz 2002). In fact, assuming a sharp dispersion law, one would be forced to conclude that the scattering matrix is trivial (a famous no-go theorem by Narnhofer *et al.* (1983)).

However, there seems to be a possibility to find an effective theory, which is much simpler and still reproduces the correct asymptotic behavior of the full theory. Disregarding low-energy excitations, Bros and Buchholz (2002) have shown that the δ -contributions in the damping factors give rise to asymptotically leading terms which have a rather simple form: they are products of the thermal correlation function of a free field and a damping factor describing the dissipative effects of the model-dependent thermal background. This result is based on the assumption that the truncated n -point functions satisfy

$$\lim_{T \rightarrow \infty} T^{3(n-1)/2 - \kappa} \mathcal{W}_\beta^{(n)}(t_1, x_1, \dots, t_n, x_n)^{\text{trunc.}} = 0 \\ \kappa > 0$$

while the δ -contribution in the damping factors exhibit, for large timelike separations T , a $T^{-3/2}$ type behavior (in $3 + 1$ spacetime dimensions).

Bros and Buchholz (2002) have shown that the asymptotically dominating parts of the correlation functions can be interpreted in terms of quasifree states acting on the algebra generated by a Hermitian field ϕ_0 satisfying the commutation relations

$$[\phi_0(t_1, x_1), \phi_0(t_2, x_2)] \\ = \Delta_{m_0}(t_1 - t_2, x_1, x_2) Z(x_1 - x_2)$$

Here Δ_{m_0} is the usual commutator function of a free scalar field of mass m_0 and Z is an operator-valued distribution commuting with ϕ_0 such that $\hat{\omega}_\beta(Z(x_1 - x_2)) = \mathcal{D}_{\beta, d}(x_1 - x_2)$. (Here $\hat{\omega}_\beta$ denotes a KMS state for the algebra generated by ϕ_0 .) Intuitively speaking, the field ϕ_0 carries an additional stochastic degree of freedom, which manifests itself in a central element that appears in the commutation relations and couples to the thermal background.

As ϕ_0 describes the interacting field asymptotically, one may expect that ϕ_0 satisfies the field

equation of the interacting field in an asymptotic sense. Buchholz and Bros (2002) have demonstrated that this assumption allows one to derive an explicit expression for the discrete part of the damping factors $\mathcal{D}_{\beta,d}(x)$ in simple models.

See also: Axiomatic Quantum Field Theory; Quantum Field Theory in Curved Spacetime; Scattering in Relativistic Quantum Field Theory: The Analytic Program; Tomita–Takesaki Modular Theory.

Further Reading

- Borchers H-J and Yngvason J (1999) Modular groups of quantum fields in thermal states. *Journal of Mathematical Physics* 40: 601–624.
- Birke L and Fröhlich J (2002) KMS, etc., *Reviews in Mathematical Physics* 14: 829–871.
- Bros J and Buchholz D (1994) Towards a relativistic KMS condition. *Nuclear Physics B* 429: 291–318.
- Bros J and Buchholz D (1996) Axiomatic analyticity properties and representations of particles in thermal quantum field theory. *Annales de l'Institut Henri Poincaré* 64: 495–521.
- Bros J and Buchholz D (1998) The unmasking of thermal Goldstone bosons. *Physical Review D* 58: 125012-1.
- Bros J and Buchholz D (2002) Asymptotic dynamics of thermal quantum fields. *Nuclear Physics B* 627: 289–310.
- Buchholz D and Junglas P (1989) On the existence of equilibrium states in local quantum field theory. *Communications in Mathematical Physics* 121: 255–270.
- Buchholz D and Ojima I (1997) Spontaneous collapse of supersymmetry. *Nuclear Physics B* 498: 228–242.
- Buchholz D, Ojima I, and Roos H (2002) Thermodynamic properties of non-equilibrium states in quantum field theory. *Annals of Physics* 297: 219–242.
- Cuniberti G, De Micheli E, and Viano GA (2001) Reconstructing the thermal green functions at real times from those at imaginary times. *Communications in Mathematical Physics* 216: 59–83.
- Derezinski J, Jaksic V, and Pillet CA (2003) Perturbations of W^* -dynamics, Liouvilleans and KMS states. *Reviews in Mathematical Physics* 15: 447–489.
- Fröhlich J (1975) The reconstruction of quantum fields from Euclidean Green's functions at arbitrary temperatures. *Helvetica Physica Acta* 48: 355–363.
- Gérard C and Jäkel C (2005) Thermal quantum fields with spatially cutoff interactions in $1+1$ space-time dimensions. *Journal of Functional Analysis* 220: 157–213.
- Gérard C and Jäkel C (2005) Thermal quantum fields without cutoffs in $1+1$ space-time dimensions. *Reviews in Mathematical Physics* 17: 113–173.
- Haag R (1992) *Local Quantum Physics*. Berlin: Springer.
- Høegh-Krohn R (1974) Relativistic quantum statistical mechanics in two-dimensional space-time. *Communications in Mathematical Physics* 38: 195–224.
- Jäkel CD (2004) The relation between KMS states for different temperatures. *Annales de l'Institut Henri Poincaré* 5: 1–30.
- Kopper C, Müller VF, and Reisz T (2001) Temperature independent renormalization of finite temperature field theory. *Annales de l'Institut Henri Poincaré* 2: 387–402.
- Kuckert B (2002) Covariant thermodynamics of quantum systems: passivity, semipassivity, and the Unruh effect. *Annals of Physics* 295: 216–229.
- Landsman NP and van Weert Ch G (1987) Real- and imaginary-time field theory at finite temperature and density. *Physics Reports* 145: 141–249.
- Narnhofer H (1994) Entropy density for relativistic quantum field theory. *Reviews in Mathematical Physics* 6: 1127–1145.
- Narnhofer H, Requardt M, and Thirring W (1983) Quasiparticles at finite temperature. *Communications in Mathematical Physics* 92: 247–268.
- Steinmann O (1995) Perturbative quantum field theory at positive temperatures: an axiomatic approach. *Communications in Mathematical Physics* 170: 405–415.

Thermohydraulics *see* Newtonian Fluids and Thermohydraulics

Toda Lattices

Y B Suris, Technische Universität München,
München, Germany

© 2006 Elsevier Ltd. All rights reserved.

Lattices, or differential–difference equations, are a special class of ordinary differential equations, with the dependent variable t playing the role of time and an infinite number of dependent variables $q_n = q_n(t)$ numbered by integer indices n , characterized by a translational invariance with respect to the shift $n \rightarrow n + 1$. Due to this property, such equations are well suited for description of processes in

translationally symmetric systems like crystals. On his search for lattice models admitting interesting explicit solutions, M Toda discovered in 1967 the lattice which nowadays carries his name:

$$\ddot{q}_n = e^{q_{n+1}-q_n} - e^{q_n-q_{n-1}} \quad [1]$$

Toda lattice is one of the most celebrated systems of mathematical physics, and a large amount of literature is devoted to it and to its various generalizations. Its most prominent property is “integrability,” so that it is amenable to a rather complete exact treatment; moreover, it can be regarded as one of the basic models, illustrating all the relevant

paradigms, notions, methods, and results of the theory of integrable systems (sometimes called the theory of solitons). One has a rare possibility to read the first-hand presentation of a large body of relevant results, including the authentic story of the original discovery, in Toda (1989).

The Infinite Toda Lattice

Model

The classical infinite Toda lattice [1] describes a one-dimensional chain of unit mass particles, each one interacting with the nearest neighbors only, q_n being the displacement of the n th particle from equilibrium. It can be treated within the Hamiltonian formalism of the classical mechanics (with some care, because of the infinite number of degrees of freedom). In this framework, the second-order Newtonian equations of motion [1] are replaced by the first-order Hamiltonian ones, for the coordinates q_n and canonically conjugate momenta p_n :

$$\dot{q}_n = p_n, \quad \dot{p}_n = e^{q_{n+1}-q_n} - e^{q_n-q_{n-1}} \tag{2}$$

The corresponding Hamilton function is

$$H_2(p, q) = \frac{1}{2} \sum_{n \in \mathbb{Z}} p_n^2 + \sum_{n \in \mathbb{Z}} (e^{q_{n+1}-q_n} - 1) \tag{3}$$

One can understand infinite sums here formally, or, alternatively, one can impose suitable boundary conditions, like $q_{n+1} - q_n \rightarrow 0, p_n \rightarrow 0$ as $|n| \rightarrow \infty$ (usually one requires decay faster than any degree of $1/|n|$).

Multisoliton Solutions

M Toda found in 1967 a number of exact traveling wave solutions of this system, including the 1-soliton solution:

$$q_n(t) = \log \frac{1 + e^{-2(\gamma_1 n - \beta_1 t + \delta_1)}}{1 + e^{-2(\gamma_1 (n-1) - \beta_1 t + \delta_1)}} \tag{4}$$

or, equivalently,

$$e^{q_{n+1}(t)-q_n(t)} = 1 + \frac{\beta_1^2}{\cosh^2(\gamma_1 n - \beta_1 t + \delta_1)} \tag{5}$$

where $\gamma_1 > 0, \beta_1 = \pm \sinh \gamma_1$, and δ_1 is an arbitrary phase. Such a soliton moves with the velocity $v_1 = \beta_1/\gamma_1$ (to the right, if $v_1 > 0$, and to the left, if $v_1 < 0$). Note that the faster the soliton is, the larger its amplitude. Multisoliton solutions were constructed in 1973 by R Hirota with the help of his ingenious “direct” (or bilinear) method. They can be written as

$$e^{q_{n+1}(t)-q_n(t)} = \frac{\tau_{n+1}(t)\tau_{n-1}(t)}{\tau_n^2(t)} \tag{6}$$

where, for an M -soliton solution, $\tau_n(t)$ can be represented through the $M \times M$ determinant depending on $2M$ parameters $z_j \in (-1, 1)$ and $c_j \in \mathbb{R}$:

$$\tau_n(t) = \det \left(\delta_{ij} + \frac{c_i(t)c_j(t)(z_i z_j)^{n+1}}{1 - z_i z_j} \right)_{1 \leq i, j \leq M} \tag{7}$$

where $c_j(t) = c_j e^{\beta_j t}, \beta_j = (1/2)(z_j^{-1} - z_j)$. If one sets $z_j = \pm e^{-\gamma_j}$ with $\gamma_j > 0$, then $\beta_j = \pm \sinh \gamma_j$, and one can show that asymptotically both for $t \rightarrow -\infty$ and for $t \rightarrow +\infty$ the solution [6] looks like the sum of well-separated solitons [4] with the velocities $v_j = \beta_j/\gamma_j$ and the respective phases $\gamma_j n - \beta_j t + \delta_j^{(\pm)}$. This is usually interpreted as a particle-like behavior of solitons. One can show that the scattering of solitons is factorized:

$$\delta_j^{(+)} - \delta_j^{(-)} = \sum_{\nu_k < \nu_j} \log \left| \frac{1 - z_j z_k}{z_j - z_k} \right| - \sum_{\nu_k > \nu_j} \log \left| \frac{1 - z_j z_k}{z_j - z_k} \right| \tag{8}$$

which means that the phase shifts of individual solitons can be interpreted as coming from the pairwise interactions only.

Integrability

The infinite Toda lattice is completely integrable in the sense of the classical Hamiltonian mechanics: it admits an infinite number of functionally independent integrals of motion in involution. This was demonstrated in 1974 by M Hénon. An instance of these higher integrals of motion is given by

$$H_3(p, q) = \frac{1}{3} \sum_{n \in \mathbb{Z}} p_n^3 + \sum_{n \in \mathbb{Z}} (p_n + p_{n+1}) e^{q_{n+1}-q_n} \tag{9}$$

Hamiltonian flows corresponding to the higher integrals of motion (usually referred to as higher Toda flows) form the “Toda lattice hierarchy.” A beautiful approach to this hierarchy is based on the Lax representation of the Toda lattice, discovered in 1974 independently by H Flaschka and S Manakov. In the variables a_n, b_n , related to q_n, p_n by

$$a_n = e^{q_{n+1}-q_n}, \quad b_n = p_n \tag{10}$$

equations of motion of the Toda lattice [2] are rewritten as

$$\dot{a}_n = a_n(b_{n+1} - b_n), \quad \dot{b}_n = a_n - a_{n-1} \tag{11}$$

It turns out that eqns [11] are equivalent to the operator equation

$$\dot{L} = [L, A_+] = [A_-, L] \tag{12}$$

where L and A_{\pm} are linear difference operators with coefficients depending on a_n, b_n :

$$L = \sum_{n \in \mathbb{Z}} b_n E_{n,n} + \sum_{n \in \mathbb{Z}} a_n E_{n,n+1} + \sum_{n \in \mathbb{Z}} E_{n+1,n} \tag{13}$$

$$\begin{aligned} A_+ &= \sum_{n \in \mathbb{Z}} b_n E_{n,n} + \sum_{n \in \mathbb{Z}} E_{n+1,n} \\ A_- &= \sum_{n \in \mathbb{Z}} a_n E_{n,n+1} \end{aligned} \tag{14}$$

Here difference operators are represented as infinite matrices, $E_{m,n}$ being the matrix with the only nonvanishing element equal to 1 in the position (m,n) . A diagonal similarity (gauge) transformation of the matrix L leads to an equivalent Lax representation of the Toda lattice:

$$\dot{L}_0 = [L_0, A_0] \tag{15}$$

with

$$L_0 = \sum_{n \in \mathbb{Z}} b_n E_{n,n} + \sum_{n \in \mathbb{Z}} a_n^{1/2} (E_{n+1,n} + E_{n,n+1}) \tag{16}$$

$$A_0 = \frac{1}{2} \sum_{n \in \mathbb{Z}} a_n^{1/2} (E_{n+1,n} - E_{n,n+1}) \tag{17}$$

Being equivalent for the Toda lattice, these two Lax representations admit nonequivalent generalizations (see below). Note that the matrices A_{\pm} in [14] may be interpreted as $A_{\pm} = \pi_{\pm}(L)$, where π_{\pm} stands for the lower-triangular, resp., strictly upper-triangular part. The commuting higher members of the Toda lattice hierarchy (enumerated by $s \in \mathbb{N}$) are characterized by the Lax equations of the form [12] with the same Lax matrix L as in [13] and with $A_{\pm} = \pi_{\pm}(L^s)$. In the Lax representation [15], the higher Toda flows are obtained by choosing $A_0 = \text{skew}(L_0^s)$, where “skew” denotes the skew-symmetric part (strictly lower-triangular part minus strictly upper-triangular part) of the symmetric matrix. The Hamilton functions of the higher flows are obtained as $H_s \sim \text{tr}(L^s) = \text{tr}(L_0^s)$.

Inverse Scattering

H Flaschka and S Manakov laid the Lax representation into the base of the application of the inverse-scattering, or inverse-spectral, transformation method (IST) to the infinite Toda lattice. It was the

first application of IST in the lattice context. The matrix L_0 in [16] is symmetric tridiagonal, which yields that the operator L_0 is second order and self-adjoint. The direct and inverse-spectral problem for $L_0 \psi = \mu \psi$ with such operators L_0 is well studied and parallel, to a large extent, to the corresponding theory for second-order differential operators. In the rapidly decaying case, the set of spectral data of the operator L_0 , allowing for a solution of the inverse problem, consists of:

1. eigenvalues $\mu_j = z_j + z_j^{-1}$ of the discrete spectrum, with $z_j \in (-1, 1)$;
2. normalizing coefficients γ_j of the corresponding eigenfunctions; and
3. reflection coefficient $r(z)$ for $|z| = 1$, characterizing the continuous spectrum $\mu = z + z^{-1} \in [-2, 2]$.

The solution of the inverse-spectral problem is given in terms of the Riemann–Hilbert problem or its variants, like the Gelfand–Levitan equation. Equation [12] means that the evolution of the operator L , induced by the evolution of $q_n(t), p_n(t)$ in virtue of the Toda lattice equations [2], is “isospectral.” More precisely, the discrete eigenvalues are integrals of motion, while the evolution of other spectral data is governed by simple linear equations:

$$\begin{aligned} z_j &= \text{const.}, \quad \gamma_j(t) = \gamma_j(0) e^{(z_j^{-1} - z_j)t} \\ r(z, t) &= r(z, 0) e^{(z^{-1} - z)t} \end{aligned} \tag{18}$$

In particular, the multisoliton solutions correspond to the reflectionless case $r(z, t) \equiv 0$. The IST solution of the initial-value problem for the infinite Toda lattice can be schematically depicted as in Figure 1.

Bi-Hamiltonian Structure

The canonical Poisson bracket for the variables q_n, p_n turns in the Flaschka–Manakov variables [10] into

$$\{b_n, a_n\}_1 = -a_n, \quad \{a_n, b_{n+1}\}_1 = -a_n \tag{19}$$

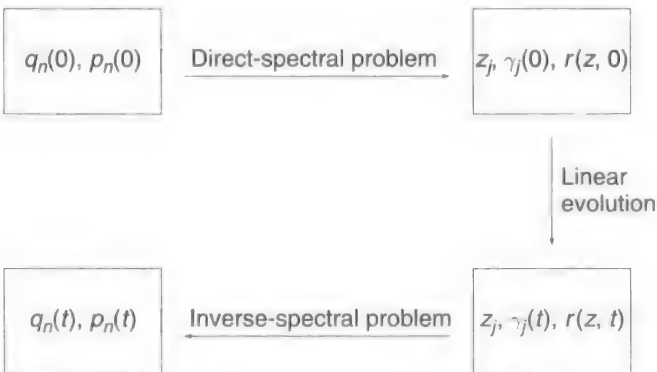


Figure 1 General scheme of the IST.

(all other brackets of the coordinate functions vanish), and the system [11] is Hamiltonian with respect to this bracket, with the Hamilton function

$$H_2 = \frac{1}{2} \sum b_n^2 + \sum a_n$$

However, one can define also a different Poisson bracket for the variables a_n, b_n :

$$\begin{aligned} \{b_n, a_n\}_2 &= -b_n a_n \\ \{a_n, b_{n+1}\}_2 &= -a_n b_{n+1} \\ \{b_n, b_{n+1}\}_2 &= -a_n \\ \{a_n, a_{n+1}\}_2 &= -a_n a_{n+1} \end{aligned} \quad [20]$$

with the following properties: it is compatible with the first one (i.e., their linear combinations are again Poisson brackets), and the system [11] is Hamiltonian with respect to this bracket, with the Hamilton function $H_1 = \sum b_n$. So, the Toda lattice in the form [11] is a bi-Hamiltonian system. This result is due to M Adler (1979). The bi-Hamiltonian property, introduced by F Magri in 1978 on the example of the Korteweg–de Vries equation, has been established since then as an alternative (and highly effective and informative) definition of integrability. Actually, the Toda lattice [11] is even tri-Hamiltonian, since there exists one more local Poisson bracket for the variables a_n, b_n with similar properties, discovered by B Kupershmidt in 1985.

Darboux–Bäcklund Transformations and Discretization

A further indispensable attribute of integrable systems are the so-called Darboux–Bäcklund transformations. For the Toda lattice they were first found by M Toda and M Wadati in 1975. A Bäcklund transformation $(q_n, p_n) \mapsto (\tilde{q}_n, \tilde{p}_n)$ with the parameter h can be written as

$$\begin{aligned} 1 + h p_n &= e^{\tilde{q}_n - q_n} + h^2 e^{q_n - \tilde{q}_{n-1}} \\ 1 + h \tilde{p}_n &= e^{\tilde{q}_n - q_n} + h^2 e^{q_{n+1} - \tilde{q}_n} \end{aligned} \quad [21]$$

This is a canonical transformation, possessing a classical generating function. These formulas can be given a fundamentally important interpretation in terms of the matrices

$$U_+ = \sum_{n \in \mathbb{Z}} e^{\tilde{q}_n - q_n} E_{n,n} + h \sum_{n \in \mathbb{Z}} E_{n+1,n} \quad [22]$$

$$U_- = I + h \sum_{n \in \mathbb{Z}} e^{q_{n+1} - \tilde{q}_n} E_{n,n+1} \quad [23]$$

The first formula in [21] is equivalent to the factorization $I + hL = U_+ U_-$, while the second one is equivalent to the factorization $I + h\tilde{L} = U_- U_+$

with the flipped factors. The Bäcklund transformation [21] serves also as an integrable discretization of the Toda flow [2] with the time step h .

Finite Open-End Toda Lattice

Model

The infinite Toda lattice [1] can be reduced to finite-dimensional systems by imposing suitable boundary conditions, different from the rapidly decaying ones. Particularly important are “open-end boundary conditions,” which correspond to placing the particles 0 and $N + 1$ at $q_0 = +\infty$ and $q_{N+1} = -\infty$, respectively. In terms of the Flaschka–Manakov variables, this means that $a_0 = a_N = 0$ and $b_0 = b_{N+1} = 0$. The Hamilton function of the resulting system with N degrees of freedom is

$$H_2(p, q) = \frac{1}{2} \sum_{n=1}^N p_n^2 + \sum_{n=1}^{N-1} e^{q_{n+1} - q_n} \quad [24]$$

This system consists of N particles subject to repulsive forces between nearest neighbors, and exhibits a scattering behavior both as $t \rightarrow -\infty$ and $t \rightarrow +\infty$. It admits a Lax representation of the same form [12] or [15] as in the infinite case, but with all the matrices being now of finite size $N \times N$, so that [13]–[14] and [16]–[17] are replaced by

$$L = \sum_{n=1}^N b_n E_{n,n} + \sum_{n=1}^{N-1} a_n E_{n,n+1} + \sum_{n=1}^{N-1} E_{n+1,n} \quad [25]$$

$$\begin{aligned} A_+ &= \sum_{n=1}^N b_n E_{n,n} + \sum_{n=1}^{N-1} E_{n+1,n} \\ A_- &= \sum_{n=1}^{N-1} a_n E_{n,n+1} \end{aligned} \quad [26]$$

and

$$L_0 = \sum_{n=1}^N b_n E_{n,n} + \sum_{n=1}^{N-1} a_n^{1/2} (E_{n+1,n} + E_{n,n+1}) \quad [27]$$

$$A_0 = \frac{1}{2} \sum_{n=1}^{N-1} a_n^{1/2} (E_{n+1,n} - E_{n,n+1}) \quad [28]$$

The qualitative behavior of the solutions is easily understood: as a consequence of repulsive interactions, the pairwise distances between particles grow infinitely, $a_n(t) = e^{q_{n+1}(t) - q_n(t)} \rightarrow 0$ as $t \rightarrow \pm\infty$, so that the matrix L_0 becomes asymptotically diagonal, with the limit velocities $b_n(\pm\infty) = \dot{q}_n(\pm\infty)$ as the diagonal entries. Due to the isospectral evolution of L_0 , these limit velocities have to coincide with the eigenvalues μ_j of L_0 , which are integrals of motion.

As $t \rightarrow -\infty$, they appear on the diagonal in the increasing order (the rightmost particle q_1 being the slowest, and the leftmost q_N being the fastest), while as $t \rightarrow +\infty$, their order on the diagonal changes to the decreasing one (the particle q_1 becoming the fastest and q_N becoming the slowest).

Moser’s Solution

Integration of this system has been first performed by J Moser in 1975. His solution can be interpreted within the general scheme of the IST (see Figure 1). The spectral data in this case consist, for example, of the eigenvalues $\mu_j (j = 1, \dots, N)$ of the matrix L_0 and the first components r_j of the corresponding orthonormal eigenvectors. The evolution of these data induced by the Toda flow [2] turns out to be simple:

$$\mu_j = \text{const.}, \quad r_j^2(t) = \frac{r_j^2(0)e^{\mu_j t}}{\sum_{i=1}^N r_i^2(0)e^{\mu_i t}} \tag{29}$$

The IST is expressed by the identity

$$\sum_{i=1}^N \frac{r_i^2}{\mu - \mu_j} = \frac{1}{\mu - b_1 - \frac{a_1}{\mu - b_2 - \dots - \frac{a_{N-1}}{\mu - b_N}}} \tag{30}$$

both parts of which represent the entry (1, 1) of the matrix $(\mu I - L_0)^{-1}$. It implies that all variables $a_n(t), b_n(t)$ are rational functions of μ_j and $e^{\mu_j t}$; in particular, one finds:

$$a_n(t) = e^{q_{n+1}(t) - q_n(t)} = \frac{\tau_{n-1}(t)\tau_{n+1}(t)}{\tau_n^2(t)} \tag{31}$$

where $\tau_n(t)$ can be represented as an $n \times n$ Hankel determinant

$$\begin{aligned} \tau_n(t) &= \det(c_{j+k}(t))_{0 \leq j, k \leq n-1} \\ c_j(t) &= \sum_{i=1}^N \mu_i^j r_i^2(t) \end{aligned} \tag{32}$$

Factorization Solution

The Lax representation [12] is a particular instance of a general construction, known under the name of Adler–Kostant–Symes (AKS) method and found around 1980. The ingredients of this construction are:

- a Lie algebra \mathfrak{g} , equipped with a nondegenerate scalar product which is used to identify \mathfrak{g} with its dual space \mathfrak{g}^* ;
- a splitting of \mathfrak{g} into a direct sum of its two subspaces \mathfrak{g}_{\pm} which are also Lie subalgebras, with $\pi_{\pm} : \mathfrak{g} \rightarrow \mathfrak{g}_{\pm}$ being the corresponding projections;

- the Lie group G of the Lie algebra \mathfrak{g} , and its Lie subgroups G_{\pm} with the Lie algebras \mathfrak{g}_{\pm} ; and
- a function $\phi : \mathfrak{g} \rightarrow \mathfrak{g}$ covariant with respect to the adjoint action of G (in the case of matrix Lie algebras and groups, one can take, e.g., $\phi(L) = L^s$).

The AKS method provides a formula for the solution of the initial-value problem for Lax equations of the form [12] with the Lax matrix $L \in \mathfrak{g}$ and $A_{\pm} = \pi_{\pm}(\phi(L))$. The solution is given by

$$L(t) = U_+^{-1}(t)L(0)U_+(t) = U_-(t)L(0)U_-^{-1}(t) \tag{33}$$

where the elements $U_{\pm}(t) \in G_{\pm}$ solve the factorization problem

$$\exp(t\phi(L(0))) = U_+(t)U_-(t) \tag{34}$$

For the open-end Toda lattice $\mathfrak{g} = \mathfrak{gl}(N)$, the Lie algebra of all $N \times N$ matrices, \mathfrak{g}_{\pm} consist of all lower-triangular, resp., strictly upper-triangular, matrices. Accordingly, $G = \text{GL}(N)$, the Lie group of all nondegenerate $N \times N$ matrices, and G_{\pm} consist of all nondegenerate lower-triangular matrices, resp., of upper-triangular matrices with units on the diagonal. The corresponding factorization problem in G is well known in the linear algebra under the name of LR factorization, and is related to the Gaussian elimination. From [33] and the well-known expression of the diagonal elements of the lower-triangular factor in the LR factorization through the minors of the factorized matrix, we find:

$$a_n(t) = \frac{\tau_{n+1}(t)\tau_{n-1}(t)}{\tau_n^2(t)} a_n(0) \tag{35}$$

where $\tau_n(t)$ is the upper-left $n \times n$ minor of the matrix $\exp(tL(0))$. If $L(t)$ is the Lax matrix along the solution of the Toda flow ($\phi(L) = L$), then the sampling of the matrix $\exp(L(t))$ at the integer times $t \in \mathbb{Z}$ coincides with the result of application of the Rutishauser’s LR algorithm to the matrix $\exp(L(0))$. The LR algorithm applied to the matrix $I + bL(0)$ is nothing other but the Bäcklund transformation [21] in the open-end situation.

Finite Periodic Toda Lattice

Model

A different reduction of the infinite Toda lattice to a finite-dimensional system appears by imposing periodic boundary conditions, $q_{n+N}(t) \equiv q_n(t)$ for all $n \in \mathbb{Z}$ (of course, such relations hold also for the Flaschka–Manakov variables a_n, b_n). The Hamilton

function of the resulting system with N degrees of freedom is

$$H_2(p, q) = \frac{1}{2} \sum_{n \in \mathbb{Z}/N\mathbb{Z}} p_n^2 + \sum_{n \in \mathbb{Z}/N\mathbb{Z}} e^{q_{n+1} - q_n} \quad [36]$$

This system consists of N particles $q_n (n = 1, \dots, N)$, and it is always assumed that $q_{N+1} \equiv q_1$ and $q_0 \equiv q_N$. Thus, the potential energy in [36] differs from the potential energy in [24] by one additional term $e^{q_1 - q_N}$. However, this modest difference leads to much more complicated dynamics of the system (quasiperiodic instead of scattering). It is convenient to replace infinite matrices in the Lax representation [12] by finite ones, of size $N \times N$, but depending on an additional parameter λ (called the spectral parameter):

$$L = \sum_{n \in \mathbb{Z}/N\mathbb{Z}} b_n E_{n,n} + \lambda^{-1} \sum_{n \in \mathbb{Z}/N\mathbb{Z}} a_n E_{n,n+1} + \lambda \sum_{n \in \mathbb{Z}/N\mathbb{Z}} E_{n+1,n} \quad [37]$$

$$A_+ = \sum_{n \in \mathbb{Z}/N\mathbb{Z}} b_n E_{n,n} + \lambda \sum_{n \in \mathbb{Z}/N\mathbb{Z}} E_{n+1,n} \quad [38]$$

$$A_- = \lambda^{-1} \sum_{n \in \mathbb{Z}/N\mathbb{Z}} a_n E_{n,n+1} \quad [39]$$

The Lax representation [12] holds identically in λ , so that the spectral parameter drops out of the equations of motion. Note that, unlike the open-end case, L is no more a tridiagonal matrix, because of the nonvanishing entries in the positions $(N, 1)$ and $(1, N)$.

Inverse-Spectral Transformation

Solution of the periodic lattice in terms of multidimensional theta functions has been given independently by E Date and S Tanaka, and by I Krichever in 1976. In this case, the set of the spectral data is more complicated; it includes:

- a hyperelliptic Riemann surface \mathcal{R} of genus $N - 1$ determined by the eigenvalues of the periodic boundary-value problem for the operator L , or, in other words, by the equation $R(\lambda, \mu) = \det(L(\lambda) - \mu I) = 0$; and
- $N - 1$ points P_k on \mathcal{R} , which correspond to the eigenvalues of L with vanishing boundary conditions.

Due to [12], the Riemann surface \mathcal{R} itself is an integral of motion, and the evolution of points P_k is such that the image of the divisor $P_1 + \dots + P_{N-1}$ under the Abel map moves along a straight line in the Jacobi variety of \mathcal{R} . Solution of the inverse-spectral problem is given in terms of

multidimensional theta-functions by formula [35] with $\tau_n(t) = \theta(nU - tV + D)$, where U, V, D are certain vectors on the Jacobian of \mathcal{R} (the first two of them depending on the spectrum \mathcal{R} only).

Loop Algebras

The periodic Toda lattice can be included into the general AKS scheme, if one interprets the Lax matrix L as an element of the loop algebra \mathfrak{g} which consists of Laurent polynomials (in λ) with coefficients from $\mathfrak{gl}(N)$, singled out by the additional condition

$$\mathfrak{g} = \{L(\lambda) \in \mathfrak{gl}(N)[\lambda, \lambda^{-1}] : \Omega L(\lambda) \Omega^{-1} = L(\omega\lambda)\}$$

where $\Omega = \text{diag}(1, \omega, \dots, \omega^{N-1})$, $\omega = \exp(2\pi i/N)$. Subalgebras \mathfrak{g}_{\pm} consist of Laurent polynomials with respect to non-negative, resp., strictly negative powers of λ . The Lie group G corresponding to the Lie algebra \mathfrak{g} consists of $\text{GL}(N)$ -valued functions $U(\lambda)$ of the complex parameter λ , regular in $\mathbb{CP}^1 \setminus \{0, \infty\}$ and satisfying $\Omega U(\lambda) \Omega^{-1} = U(\omega\lambda)$. Its subgroups G_{\pm} corresponding to the Lie algebras \mathfrak{g}_{\pm} are singled out by the following conditions: elements of G_+ are regular in the neighborhood of $\lambda = 0$, while elements of G_- are regular in the neighborhood of $\lambda = \infty$ and take at $\lambda = \infty$ the value I . The corresponding factorization is called the generalized LR factorization. As opposed to the open-end case, finding such a factorization is a problem of the Riemann–Hilbert type which is solved in terms of algebraic geometry and theta-functions rather than in terms of linear algebra and exponential functions. This approach to the periodic Toda lattice is due to Reyman and Semenov-Tian-Shansky (1979) and, independently, to M Adler and P van Moerbeke (1980).

Generalizations: Lie-Algebraic Systems

The AKS interpretation of the finite Toda lattices leads directly to their generalizations by replacing the algebra $\mathfrak{gl}(N)$, resp., the loop algebra over $\mathfrak{gl}(N)$, by simple Lie algebras, resp. affine Lie algebras. These generalized Toda systems were introduced in 1976 by O Bogoyavlensky and solved in 1979 independently by M Olshanetsky, A Perelomov, and by B Kostant.

Simple Lie Algebras

Let \mathfrak{g} be a simple Lie algebra (complex or real split), and \mathfrak{h} its Cartan subalgebra. Let further $\Delta = \Delta_+ \cup \Delta_-$ be the root system of \mathfrak{g} , decomposed into the sets of positive roots Δ_+ and the set of negative roots Δ_- . One has a direct vector space $\mathfrak{g} = \mathfrak{g}_+ \oplus \mathfrak{g}_-$, where \mathfrak{g}_+ is spanned by the root spaces for positive roots and by \mathfrak{h} ,

while \mathfrak{g}_- is spanned by the root spaces for negative roots (Borel decomposition). For $\alpha \in \Delta$ let E_α be a corresponding root vector. So, $[H, E_\alpha] = \alpha(H)E_\alpha$ for all $H \in \mathfrak{h}$. The root $\alpha \in \mathfrak{h}^*$ may be identified with $H_\alpha \in \mathfrak{h}$ defined by $\langle H_\alpha, H \rangle = \alpha(H)$ for all $H \in \mathfrak{h}$. It is easy to deduce that $[E_\alpha, E_{-\alpha}] = c_\alpha H_\alpha$, where $c_\alpha = \langle E_\alpha, E_{-\alpha} \rangle$. The system of simple roots will be denoted by $\Phi \subset \Delta_+$.

The generalized Toda lattice for the Lie algebra \mathfrak{g} is the following system of differential equations on $\mathfrak{h} \times \mathfrak{h}$:

$$\begin{aligned} \dot{Q} &= P \\ \dot{P} &= - \sum_{\alpha \in \Phi} e^{\alpha(Q)} [E_\alpha, E_{-\alpha}] = - \sum_{\alpha \in \Phi} c_\alpha e^{\alpha(Q)} H_\alpha \end{aligned} \quad [40]$$

This system can be given a Hamiltonian formulation, with the Hamilton function

$$H_g = \frac{1}{2} \langle P, P \rangle + \sum_{\alpha \in \Phi} c_\alpha e^{\alpha(Q)} \quad [41]$$

It is completely integrable, and has a Lax representation [12] with

$$L = P + \sum_{\alpha \in \Phi} E_\alpha + \sum_{\alpha \in \Phi} e^{\alpha(Q)} E_{-\alpha} \quad [42]$$

$$A_+ = P + \sum_{\alpha \in \Phi} E_\alpha, \quad A_- = \sum_{\alpha \in \Phi} e^{\alpha(Q)} E_{-\alpha} \quad [43]$$

The usual open-end Toda lattice corresponds to the algebra $\mathfrak{sl}(N)$ (series A_{N-1}), so that the Hamilton function [24] can be denoted by $H_{A_{N-1}}$. The Hamilton functions of the generalized lattices corresponding to other classical algebras $\mathfrak{so}(2N+1)$ (series B_N), $\mathfrak{sp}(N)$ (series C_N), and $\mathfrak{so}(2N)$ (series D_N) can be written in the canonically conjugate variables $q_n, p_n (n=1, \dots, N)$ as

$$H_g(p, q) = H_{A_{N-1}}(p, q) + \begin{cases} e^{-q_N}, & \mathfrak{g} = B_N \\ e^{-2q_N}, & \mathfrak{g} = C_N \\ e^{-q_N - q_{N-1}}, & \mathfrak{g} = D_N \end{cases} \quad [44]$$

Affine Lie Algebras

Turning to the generalizations of the periodic Toda lattice, let θ be a Coxeter automorphism of a simple complex algebra \mathfrak{g} , the order of θ being m . Introduce the loop algebra \mathfrak{g} as the Lie algebra of Laurent polynomials

$$\mathfrak{g} = \{L(\lambda) \in \mathfrak{g}[\lambda, \lambda^{-1}] : \theta(L(\lambda)) = L(\omega\lambda)\}$$

where $\omega = \exp(2\pi i/m)$. Denote by \mathfrak{g}_j the eigenspaces of θ corresponding to the eigenvalues $\omega^j (j \in \mathbb{Z}/m\mathbb{Z})$. Set $\mathfrak{a} = \mathfrak{g}_0$, and let s denote the dimension of \mathfrak{a} . By definition of the Coxeter automorphism, \mathfrak{a} is an

abelian subalgebra of \mathfrak{g} . Denote by Ψ the set of $\alpha \in \mathfrak{a}^*$ for which there exist nonzero elements $E_\alpha \in \mathfrak{g}_1$ with $[H, E_\alpha] = \alpha(H)E_\alpha$ for all $H \in \mathfrak{a}$. The elements $E_{-\alpha} \in \mathfrak{g}_{-1}$ are defined similarly. It can be shown that Ψ contains $s+1$ elements, so that between them there exists exactly one linear relation. The elements of Ψ are called simple weights of the loop algebra \mathfrak{g} . The Lie algebra \mathfrak{g} is a direct sum of its two subspaces \mathfrak{g}_\pm consisting of Laurent polynomials with non-negative, resp., with strictly negative powers of λ ; these subspaces are also Lie subalgebras.

Now the generalized Toda lattice related to the loop algebra \mathfrak{g} can be introduced as the system of differential equations on $\mathfrak{a} \times \mathfrak{a}$, which looks formally exactly as [40], and has the Hamilton function which looks exactly as [41], but with the set of simple roots Φ of \mathfrak{g} being replaced by the set of simple weights Ψ of \mathfrak{g} . The matrices participating in the Lax representation [12] belong now to the loop algebra \mathfrak{g} :

$$L(\lambda) = P + \lambda \sum_{\alpha \in \Psi} E_\alpha + \lambda^{-1} \sum_{\alpha \in \Psi} e^{\alpha(Q)} E_{-\alpha} \quad [45]$$

$$\begin{aligned} A_+(\lambda) &= P + \lambda \sum_{\alpha \in \Psi} E_\alpha \\ A_-(\lambda) &= \lambda^{-1} \sum_{\alpha \in \Psi} e^{\alpha(Q)} E_{-\alpha} \end{aligned} \quad [46]$$

For the classical series of loop algebras, the Hamilton functions H_g in the canonically conjugate variables $q_n, p_n (n=1, \dots, N)$ can be presented as

$$H_g(p, q) = H_{A_{N-1}}(p, q) + \begin{cases} e^{-q_N} + e^{q_1 + q_2}, & \mathfrak{g} = B_N^{(1)} \\ e^{-2q_N} + e^{2q_1}, & \mathfrak{g} = C_N^{(1)} \\ e^{-q_N - q_{N-1}} + e^{q_1 + q_2}, & \mathfrak{g} = D_N^{(1)} \\ e^{-2q_N} + e^{q_1 + q_2}, & \mathfrak{g} = A_{2N-1}^{(2)} \\ e^{-q_N} + e^{2q_1}, & \mathfrak{g} = A_{2N}^{(2)} \\ e^{-q_N} + e^{q_1}, & \mathfrak{g} = D_{N+1}^{(2)} \end{cases} \quad [47]$$

Actually, one can find even more general integrable systems of the Toda type: one can add to $H_{A_{N-1}}(p, q)$ any of the two potentials $e^{-q_N - q_{N-1}}$ or $\alpha e^{-q_N} + \beta e^{-2q_N}$ on one end combined with any of the two potentials $e^{q_1 + q_2}$ or $\gamma e^{q_1} + \delta e^{2q_1}$ on the other end, where $\alpha, \beta, \gamma, \delta$ are arbitrary constants. This result is due to E Sklyanin (1987).

Generalizations: Lattices with Nearest-Neighbor Interactions

There exist further integrable lattice systems with the nearest-neighbor interaction apart from the classical exponential Toda lattice [1]. Those of the

type $\ddot{q}_n = r(\dot{q}_n)(g(q_{n+1} - q_n) - g(q_n - q_{n-1}))$ have been classified by R Yamilov in 1982, and the list contains, apart from the usual Toda lattice [1], the following ones:

$$\ddot{q}_n = \dot{q}_n(e^{q_{n+1}-q_n} - e^{q_n-q_{n-1}}) \quad [48]$$

$$\ddot{q}_n = \dot{q}_n(q_{n+1} - 2q_n + q_{n-1}) \quad [49]$$

$$\ddot{q}_n = -(\dot{q}_n^2 - \nu^2) \left(\frac{1}{q_{n+1} - q_n} - \frac{1}{q_n - q_{n-1}} \right) \quad [50]$$

$$\ddot{q}_n = -(\dot{q}_n^2 - \nu^2)(\coth(q_{n+1} - q_n) - \coth(q_n - q_{n-1})) \quad [51]$$

Equations [48] are known as the “modified Toda lattice.” Equations [49] describe the “dual Toda lattice” which was instrumental in the original discovery by Toda (see Toda (1989)). All systems [49]–[51] can be obtained from [11] via suitable parametrizations of the variables a_n, b_n by canonically conjugate ones q_n, p_n , similar to [10] for [1], see Suris (2003).

A remarkable discovery of the integrable relativistic Toda lattice is due to S Ruijsenaars (1990). This lattice with the equations of motion

$$\ddot{q}_n = (1 + \alpha\dot{q}_n) \left((1 + \alpha\dot{q}_{n+1}) \frac{e^{q_{n+1}-q_n}}{1 + \alpha^2 e^{q_{n+1}-q_n}} - (1 + \alpha\dot{q}_{n-1}) \frac{e^{q_n-q_{n-1}}}{1 + \alpha^2 e^{q_n-q_{n-1}}} \right) \quad [52]$$

can be considered as the perturbation of the usual Toda lattice with the small parameter α (the inverse speed of light).

A class of integrable lattice systems of the relativistic Toda type $\ddot{q}_n = r(\dot{q}_n)(\dot{q}_{n+1}f(q_{n+1} - q_n) - \dot{q}_{n-1}f(q_n - q_{n-1}) + g(q_{n+1} - q_n) - g(q_n - q_{n-1}))$ is richer than that of the Toda type, and has been isolated by Yu B Suris and by V Adler and A Shabat in 1997. The list contains, apart from the relativistic Toda lattice [52], two more α -perturbations of the usual Toda lattice [1]:

$$\ddot{q}_n = (1 + \alpha\dot{q}_{n+1})e^{q_{n+1}-q_n} - (1 + \alpha\dot{q}_{n-1})e^{q_n-q_{n-1}} - \alpha^2(e^{2(q_{n+1}-q_n)} - e^{2(q_n-q_{n-1})}) \quad [53]$$

$$\ddot{q}_n = (1 - \alpha\dot{q}_n)^2 \left((1 - \alpha\dot{q}_{n+1}) e^{q_{n+1}-q_n} - (1 - \alpha\dot{q}_{n-1}) e^{q_n-q_{n-1}} \right) \quad [54]$$

two α -perturbations of the modified Toda lattice [48]:

$$\ddot{q}_k = \dot{q}_n \left(e^{q_{n+1}-q_n} - e^{q_n-q_{n-1}} + \alpha\dot{q}_{n+1} \frac{e^{q_{n+1}-q_n}}{1 + \alpha e^{q_{n+1}-q_n}} - \alpha\dot{q}_{n-1} \frac{e^{q_n-q_{n-1}}}{1 + \alpha e^{q_n-q_{n-1}}} \right) \quad [55]$$

$$\ddot{q}_n = \dot{q}_n(1 - \alpha\dot{q}_n) \left((1 - \alpha\dot{q}_{n+1}) \frac{e^{q_{n+1}-q_n}}{1 + \alpha e^{q_{n+1}-q_n}} - (1 - \alpha\dot{q}_{n-1}) \frac{e^{q_n-q_{n-1}}}{1 + \alpha e^{q_n-q_{n-1}}} \right) \quad [56]$$

two α -perturbations of the dual Toda lattice [49]:

$$\ddot{q}_n = \dot{q}_n(q_{n+1} - 2q_n + q_{n-1}) + \frac{\alpha\dot{q}_{n+1}\dot{q}_n}{1 + \alpha(q_{n+1} - q_n)} - \frac{\alpha\dot{q}_n\dot{q}_{n-1}}{1 + \alpha(q_n - q_{n-1})} \quad [57]$$

$$\ddot{q}_n = \dot{q}_n(1 + \alpha^2\dot{q}_n) \left(\frac{q_{n+1} - q_n - \alpha\dot{q}_{n+1}}{1 + \alpha(q_{n+1} - q_n)} - \frac{q_n - q_{n-1} - \alpha\dot{q}_{n-1}}{1 + \alpha(q_n - q_{n-1})} \right) \quad [58]$$

and one α -perturbation of each of the systems [50] and [51]:

$$\ddot{q}_n = -(\dot{q}_n^2 - \nu^2) \left(\frac{q_{n+1} - q_n - \alpha\dot{q}_{n+1}}{(q_{n+1} - q_n)^2 - (\nu\alpha)^2} - \frac{q_n - q_{n-1} - \alpha\dot{q}_{n-1}}{(q_n - q_{n-1})^2 - (\nu\alpha)^2} \right) \quad [59]$$

$$\ddot{q}_n = -\frac{1}{2}(\dot{q}_n^2 - \nu^2) \times \left(\frac{\sinh 2(q_{n+1} - q_n) - \nu^{-1} \sinh(2\nu\alpha)\dot{q}_{n+1}}{\sinh^2(q_{n+1} - q_n) - \sinh^2(\nu\alpha)} - \frac{\sinh 2(q_n - q_{n-1}) - \nu^{-1} \sinh(2\nu\alpha)\dot{q}_{n-1}}{\sinh^2(q_n - q_{n-1}) - \sinh^2(\nu\alpha)} \right) \quad [60]$$

A detailed study of all these systems, their interrelations, and time discretizations can be found in Suris (2003).

There exist also lattices with more complicated nearest-neighbor interactions, involving elliptic functions. They were discovered by A Shabat and R Yamilov (1990), and by I Krichever (2000). For example, the nonrelativistic elliptic Toda lattice is governed by the equations

$$\ddot{q}_n = (\dot{q}_n^2 - 1)(V(q_n, q_{n+1}) + V(q_n, q_{n-1})) \quad [61]$$

where $V(q, q') = \zeta(q + q') + \zeta(q - q') - \zeta(2q)$ is an elliptic function in both arguments q, q' (here $\zeta(q)$ is the Weierstrass ζ -function).

Further Developments and Generalizations

Sato's Theory

Formulas [6], [31], and [35] have the same structure, with the case-dependent functions $\tau_n(t)$ given by the determinants [7] for the multisoliton solution in the

infinite case, by the Hankel determinants [32] or by the minors of the matrix $\exp(L(0))$ in the open case, and by the multidimensional theta functions in the periodic case. All these seemingly different objects are actually particular cases of a beautiful construction due to M Sato (1981), developed by E Date, M Jimbo, M Kashiwara, T Miwa (1981–83), and by G Segal and G Wilson (1985), which provides one of the major unifying schemes for the theory of integrable systems. In this construction, integrable systems are interpreted as simple dynamical systems on an infinite-dimensional Grassmannian. The τ -function (first invented by R Hirota in 1971) receives in this theory a representation-theoretical interpretation in terms of the determinant bundle over the Grassmannian.

Band Matrices

The Lax matrices [13] and [16] in the Manakov–Flaschka variables can be easily generalized: in the symmetric matrix L_0 one can admit nonvanishing elements in the band of the width $2s + 1 > 3$ around the main diagonal, in the Heisenberg matrix L one can admit more nonvanishing diagonals in the upper-triangle part. A systematic presentation of a large body of relevant results is given in Kupershmidt (1985). In the setting of finite lattices, the integrability of such systems becomes a non-trivial problem (as opposed to the tridiagonal situation), because the number of independent conjugation-invariant functions $\text{tr}(L^s)$ becomes less than the number of degrees of freedom. An effective approach to this problem based on the semi-invariant functions has been found by P Deift, L-Ch Li, T Nanda, and C Tomei in 1986.

Two-Dimensional Toda Lattices

Up to now, we considered integrable lattices with one continuous and one discrete independent variables. This allows for a further generalization. Integrable systems with two continuous and one discrete independent variables are well known and widely used as models of the field theory. For instance, the Toda field theory deals with the system

$$(q_n)_{xy} = e^{q_{n+1}-q_n} - e^{q_n-q_{n-1}} \quad [62]$$

introduced in the soliton theory by A Mikhailov in 1979. This two-dimensional system admits all possible kinds of reductions and generalizations mentioned above for the usual Toda lattice. In particular, the periodic two-dimensional Toda lattice is referred to as the affine Toda field theory (with the prominent example of the sine-Gordon field which corresponds to the period 2). Later, it was realized that the equivalent equation $(\log v_n)_{xy} = v_{n+1} - 2v_n + v_{n-1}$,

which is obtained from [62] by setting $v_n = \exp(q_{n+1} - q_n)$, already appeared in studies by G Darboux in the 1880s, as the equation satisfied by the Laplace invariants of the chain of Laplace transformations of a given conjugate net. This relation to the classical differential geometry was extensively studied by G Darboux, G Tzitzéica, and others long before the advent of the theory of integrable systems. Another link to the differential geometry is a more recent observation, and relates the two-dimensional Toda lattice, with the d'Alembert operator $(\cdot)_{xy}$ on the left-hand side of [62] replaced by the Laplace operator $(\cdot)_{zz}$, to harmonic maps. For instance, the sinh-Gordon equation $u_{z\bar{z}} = \sinh u$ governs harmonic maps from \mathbb{C} into the unit sphere S^2 , which can be interpreted also as Gauss maps of the constant mean curvature surfaces in \mathbb{R}^3 . A review of this topic can be found in Guest (1997).

Discretization of Toda lattices, nonabelian Toda Lattices, quantization of Toda lattices, dispersionless limit of Toda lattices, etc., are only some of the further relevant topics, which cannot be discussed in any detail in the restricted frame of this article, and the same holds, unfortunately, for such fascinating applications of the Toda lattice as the Frobenius manifolds, Laplacian growth problem, quantum cohomology, random matrix theory, two-dimensional gravity, etc.

See also: Bäcklund Transformations; Bi-Hamiltonian Methods in Soliton Theory; Classical r -Matrices, Lie Bialgebras, and Poisson Lie Groups; Current Algebra; Dynamical Systems and Thermodynamics; Functional Equations and Integrable Systems; Integrable Discrete Systems; Integrable Systems and Discrete Geometry; Integrable Systems and the Inverse Scattering Method; Integrable Systems: Overview; Lie Groups: General Theory; Multi-Hamiltonian Systems; Quantum Calogero–Moser Systems; Separation of Variables for Differential Equations; Solitons and Kac–Moody Lie Algebras; WDVV Equations and Frobenius Manifolds.

Further Reading

- Adler M (1979) On a trace functional for formal pseudo-differential operators and the symplectic structure for Korteweg–de Vries type equations. *Invent. Mathematics* 50: 219–248.
- Adler VE and Shabat AB (1997) On a class of Toda chains. *Teor. Mat. Phys.* 111: 323–334 (in Russian; English translation: *Theor. Math. Phys.* 111: 647–657); Generalized Legendre transformations. *Teor. Mat. Phys.* 112: 179–194 (in Russian; English translation: *Theor. Math. Phys.* 112: 935–948).
- Adler M and van Moerbeke P (1980) Completely integrable systems, Kac–Moody algebras and curves; *Advances in Mathematics* 38: 267–317. Linearization of Hamiltonian systems, Jacobi varieties and representation theory. *Advances in Mathematics* 38: 318–379.

Bogoyavlensky OI (1976) On perturbations of the periodic Toda lattice. *Communications in Mathematical Physics* 51: 201–209.

Date F, Jimbo M, and Miwa T (1982–83) Method for generating discrete soliton equations. I–V. *Journal of the Physical Society of Japan* 51: 4116–4124, 4125–4131; 52: 388–393, 761–765, 766–771.

Date E and Tanaka S (1976) Analogue of inverse scattering theory for the discrete Hill's equation and exact solutions for the periodic Toda lattice. *Progress in Theoretical Physics* 55: 457–465.

Deift P, Li L-C, Nanda T, and Tomei C (1986) The Toda flow on a generic orbit is integrable. *Communications in Pure and Applied Mathematics* 39: 183–232.

Flaschka H (1974) On the Toda lattice II. Inverse scattering solution. *Progress in Theoretical Physics* 51: 703–716.

Guest MA (1997) *Harmonic Maps, Loop Groups, and Integrable Systems*. Cambridge: Cambridge University Press.

Hénon M (1974) Integrals of the Toda lattice. *Physical Review B* 9: 1921–1923.

Krichever IM (1978) Algebraic curves and nonlinear difference equations. *Uspekhi Mat. Nauk* 33: 215–216 (in Russian).

Krichever I (2000) Elliptic analog of the Toda Lattice. *Int. Math. Res. Notes* 8: 383–412.

Kupershmidt BA (1985) Discrete Lax equations and differential-difference calculus. *Asterisque* 123: 212.

Magri F (1978) A simple model of the integrable Hamiltonian equation. *Journal of Mathematical Physics* 19: 1156–1162.

Manakov SV (1974) On the complete integrability and stochasticization in discrete dynamical systems. *Zh. Exp. Theoretical Physics* 67: 543–555 (in Russian; English translation: *Soviet Physics JETP* 40: 269–274).

Mikhailov AV (1979) Integrability of a two-dimensional generalization of the Toda chain. *Pis'ma Zh. Eksp. Teor. Fiz.* 30: 443–448 (in Russian; English translation: *JETP Letters* 30: 414–418).

Moser J (1975) Finitely many mass points on the line under the influence of an exponential potential – an integrable system. *Lecture Notes Physics* 38: 467–497.

Olshanetsky MA and Perelomov AM (1979) Explicit solutions of classical generalized Toda models. *Invent. Math.* 54: 261–269.

Reyman AG and Semenov-Tian-Shansky MA (1979–81) Reduction of Hamiltonian systems, affine Lie algebras and Lax equations. I, II. *Invent. Math.* 54: 81–100, 63: 423–432.

Reyman AG and Semenov-Tian-Shansky MA (1994) Group theoretical methods in the theory of finite dimensional integrable systems. In: Arnold VI and Novikov SP (eds.) *Encyclopaedia of Mathematical Science*, vol. 16. *Dynamical Systems VII*, pp. 116–225. Berlin: Springer.

Ruijsenaars SNM (1990) Relativistic Toda systems. *Communications in Mathematical Physics* 133: 217–247.

Sato M and Sato Y (1983) Soliton equations as dynamical systems on infinite-dimensional Grassmann manifold. In: *Nonlinear Partial Differential Equations in Applied Science (Tokyo, 1982)*, pp. 259–271. North-Holland: Amsterdam.

Segal G and Wilson G (1985) Loop groups and equations of KdV type. *Inst. Hautes Études Sci. Publ. Math.* 61: 5–65.

Shabat AB and Yamilov RI (1990) Symmetries on nonlinear chains. *Algebra i Analiz* 2: 183–208 (in Russian; English translation: *Leningrad Mathematical Journal* 2: 377–400).

Sklyanin EK (1987) Boundary conditions for integrable equations. *Funkts. Anal. Prilozh.* 21: 86–87 (in Russian; English translation: *Funct. Anal. Appl.* 21: 164–166).

Suris YB (1997) New integrable systems related to the relativistic Toda lattice. *Journal of Physics A: Math. and Gen.* 30: 1745–1761.

Suris YuB (2003) *The Problem of Integrable Discretization: Hamiltonian Approach*. Basel: Birkhäuser.

Toda M (1967) Vibration of a chain with nonlinear interaction. *Journal of Physical Society Japan* 22: 431–436.

Toda M (1989) *Theory of Nonlinear Lattices*. Berlin: Springer.

Toda M and Wadati M (1975) A canonical transformation for the exponential lattice. *Journal of the Physical Society of Japan* 39: 1204–1211.

Yamilov RI (1982) On the classification of discrete equations. In: *Integrable Systems*, Ufa, 95–114 (in Russian).

Toeplitz Determinants and Statistical Mechanics

E L Basor, California Polytechnic State University, San Luis Obispo, CA, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

A finite Toeplitz matrix is an $n \times n$ matrix with the following structure:

$$\begin{pmatrix} a_0 & a_{-1} & a_{-2} & \cdots & a_{-n+1} \\ a_1 & a_0 & a_{-1} & \cdots & a_{-n+2} \\ a_2 & a_1 & a_0 & \cdots & a_{-n+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n-1} & a_{n-2} & a_{n-3} & \cdots & a_0 \end{pmatrix} \tag{1}$$

The entries depend on the difference $i - j$ and hence they are constant down all the diagonals. There are two cases when the determinant is easy to compute. One is when the matrix is upper- or lower-triangular

and the determinant is a_0^n . The other case is when the matrix is of the form

$$\begin{pmatrix} a_0 & a_{n-1} & a_{n-2} & \cdots & a_1 \\ a_1 & a_0 & a_{n-1} & \cdots & a_2 \\ a_2 & a_1 & a_0 & \cdots & a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n-1} & a_{n-2} & a_{n-3} & \cdots & a_0 \end{pmatrix} \tag{2}$$

In this latter case, the matrix is called a circulant matrix and the eigenvalues are given by the formula

$$\psi_n(e^{i2\pi k/n}), \quad 0 \leq k \leq n - 1$$

where

$$\psi_n(e^{i\theta}) = \sum_{j=0}^{n-1} a_j e^{ij\theta}$$

The corresponding eigenvector for eigenvalue $\psi_n(e^{i2\pi k/n})$ is

$$[1, e^{i2\pi k/n}, \dots, e^{i2\pi k(n-1)/n}]$$

This can be verified by direct computation. The role of circulant matrices will not be emphasized in this article, although they are used in the computation of the generating function for certain dimer configurations and also in applications using the discrete Fourier transform.

The most common way to generate a finite Toeplitz matrix is with the Fourier coefficients of an integrable function. Let $\phi: T \rightarrow \mathbb{C}$ be a function defined on the unit circle with Fourier coefficients

$$\phi_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi(e^{i\theta}) e^{-ik\theta} d\theta \quad [3]$$

We define $T_n(\phi)$ to be the Toeplitz matrix:

$$T_n(\phi) = (\phi_{i-j})_{i,j=0}^{n-1}$$

A basic problem that in large part has been motivated by statistical mechanics is to determine the behavior of the asymptotics of the determinant of $T_n(\phi)$ as $n \rightarrow \infty$. The determinant will be referred to as $D_n(\phi)$, where ϕ is called the generating function of the determinant. If the generating function has the property that its Fourier coefficients vanish for negative index (positive index) then the corresponding matrix is lower-triangular (upper-triangular) and hence the determinant is ϕ_0^n . For other cases, the determinant is not easy to determine and requires additional mathematical machinery.

Some of the primary motivation to study the determinant of these matrices comes from the two-dimensional Ising model. We consider the Onsager lattice in the absence of a magnetic field with sites labeled by

$$(i, j), \quad 0 \leq i \leq M, \quad 0 \leq j \leq N$$

and with a value $\sigma_{i,j} = \pm 1$ assigned to each site. In the Ising model, $\sigma_{i,j}$ signifies the state of the spin at the site (i, j) . To each possible configuration of spins, we define an energy

$$\mathcal{E}(\sigma) = -E_1 \sum_{i,j} \sigma_{i,j} \sigma_{i+1,j} - E_2 \sum_{i,j} \sigma_{i,j} \sigma_{i,j+1}$$

Let

$$Z = \sum_{\sigma} e^{-\beta \mathcal{E}(\sigma)}$$

be the partition function. Then the probability of a given configuration is

$$\frac{1}{Z} e^{-\beta \mathcal{E}(\sigma)}$$

Here E_1, E_2 , and $\beta = 1/kT$ are, without loss of generality, assumed to be positive constants, T is the temperature, and k is the Boltzmann constant. If X is a random variable defined on the space of configurations, the expectation is given by

$$E(X) = \frac{1}{Z} \sum_{\sigma=\pm 1} X(\sigma) e^{-\beta \mathcal{E}(\sigma)}$$

Let n be fixed for the moment and assume toroidal boundary conditions for the lattice and then let $N, M \rightarrow \infty$. It is known that the random variable

$$X(\sigma) = \sigma_{0,0} \sigma_{0,n}$$

has expectation $\langle \sigma_{0,0} \sigma_{0,n} \rangle$ given by $D_n(\phi)$, where

$$\phi(e^{i\theta}) = \left(\frac{(1 - \alpha_1 e^{i\theta})(1 - \alpha_2 e^{-i\theta})}{(1 - \alpha_1 e^{-i\theta})(1 - \alpha_2 e^{i\theta})} \right)^{1/2}$$

$$\alpha_1 = z_1 \left(\frac{1 - z_2}{1 + z_2} \right), \quad \alpha_2 = z_1^{-1} \left(\frac{1 - z_2}{1 + z_2} \right)$$

and

$$z_1 = \tanh \beta E_1, \quad z_2 = \tanh \beta E_2$$

The square root is taken so that $\phi(e^{i\pi}) = 1$. This formula was first stated by Onsager and later verified in a difficult computation by Montroll, Potts, and Ward.

The spontaneous magnetization M for the Ising model is defined by

$$M^2 = \lim_{n \rightarrow \infty} \langle \sigma_{0,0} \sigma_{0,n} \rangle = \lim_{n \rightarrow \infty} D_n(\phi)$$

Note that it is the square root of the correlation between two distant sites. Hence, the asymptotics of the Toeplitz determinants will determine whether the magnetization is positive or tends to zero as $n \rightarrow \infty$.

Strong Szegő Limit Theorem

To determine the behavior of the determinants, we need to analyze the generating function ϕ . Let us first consider the case where $\alpha_2 < 1$. (It is always the case that $0 < \alpha_1 < 1$.) This generating function is differentiable, nonzero and has winding number zero, and it is for functions of this type that a second-order expansion of the Toeplitz determinants can be described. The expansion first formulated by Szegő, in response to the question concerning the

spontaneous magnetization, is called the “strong Szegő limit theorem.”

Before proving the Szegő theorem, it should be remarked that we can view the finite Toeplitz matrix as a truncation of an infinite array,

$$\begin{pmatrix} \phi_0 & \phi_{-1} & \phi_{-2} & \ddots \\ \phi_1 & \phi_0 & \phi_{-1} & \ddots \\ \phi_2 & \phi_1 & \phi_0 & \ddots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad [4]$$

The above infinite array is the matrix representation for the Toeplitz operator

$$T(\phi) : H^2 \rightarrow H^2$$

defined by

$$T(\phi)f = P(\phi f)$$

where H^2 is the Hardy space

$$\{f \in L^2(T) \mid f_k = 0, k < 0\}$$

the function $\phi \in L^\infty(T)$, and P is the orthogonal projection of $L^2(T)$ onto H^2 . The matrix representation given in [4] is with respect to the Hilbert space basis of H^2 ,

$$\{e^{ik\theta} \mid 0 \leq k < \infty\}$$

and ϕ is called the symbol of the operator. Now define $P_n : H^2 \rightarrow H^2$ by

$$P_n(f_0, f_1, f_2, \dots) = (f_0, f_1, f_2, \dots, f_{n-1}, 0, 0, \dots)$$

The finite Toeplitz matrix can be thought of as the upper-left corner of the array given in [4] or as $P_n T(\phi) P_n$.

To prove the Strong Szegő limit theorem, we introduce the Banach algebra \mathcal{B} of bounded functions f satisfying $\sum_{k=-\infty}^{\infty} \|f_k\|^2 < \infty$.

Theorem 1 (Strong Szegő limit theorem). *Assume $\phi = \phi_- \phi_+$, where ϕ_\pm have logarithms in \mathcal{B} . Suppose $\log \phi_-$, $\log \phi_+ \in H^2$. Then*

$$\lim_{n \rightarrow \infty} D_n(\phi) / G(\phi)^n = E(\phi) = \exp \left(\sum_{k=1}^{\infty} k s_k s_{-k} \right)$$

where $G(\phi) = \exp((\log \phi)_0)$ and $s_k = \log \phi_k$.

Since \mathcal{B} is a Banach algebra, it follows that if $\log \phi_\pm$ belong to \mathcal{B} so do

$$\phi_-, \phi_+, \phi_+^{-1}, \phi_-^{-1}, \phi, \phi^{-1}$$

and hence they are bounded. Since ϕ_+ is in H^2 as well, its Fourier coefficients vanish for negative index and the Toeplitz operator has a corresponding infinite array that is lower-triangular. The Fourier coefficients vanish for positive index for ϕ_- and

hence the infinite array is upper triangular. From this, it follows that

$$T(\phi_+)T(\phi_+^{-1}) = T(\phi_-^{-1})T(\phi_-) = I \quad [5]$$

$$T(\phi_-)T(\phi_+) = T(\phi) \quad [6]$$

and

$$P_n T(\phi_+) = P_n T(\phi_+) P_n \quad [7]$$

$$P_n T(\phi_-) P_n = T(\phi_-) P_n$$

This yields

$$D_n(\phi) = \det T_n(\phi) = \det P_n T(\phi) P_n \quad [8]$$

$$= \det P_n T(\phi_+) T(\phi_+^{-1}) T(\phi) T(\phi_-^{-1}) T(\phi_-) P_n \quad [9]$$

$$= \det P_n T(\phi_+) P_n T(\phi_+^{-1}) T(\phi) T(\phi_-^{-1}) P_n T(\phi_-) P_n \quad [10]$$

$$z = \det P_n T(\phi_+) P_n \det (P_n T(\phi_+^{-1}) T(\phi) T(\phi_-^{-1}) P_n) \times \det P_n T(\phi_-) P_n \quad [11]$$

The determinants of the right-hand side and the left-hand side of the above expression are $((\phi_\mp)_0)^n$, respectively. Now given the Banach algebra conditions imposed on the symbol ϕ , it follows that the operator

$$T(\phi_+^{-1})T(\phi)T(\phi_-^{-1})$$

is of the form $I + K$, where K is trace class. Hence, the eigenvalues λ_i of K satisfy

$$\sum |\lambda_i| < \infty$$

and the infinite (Fredholm) determinant of $I + K$ is defined. To verify the claim that the operator

$$T(\phi_+^{-1})T(\phi)T(\phi_-^{-1}) = T(\phi_+^{-1})T(\phi_-)T(\phi_+)T(\phi_-^{-1})$$

is I plus a trace class operator, we use the identity

$$T(fg) - T(f)T(g) = H(f)H(\tilde{g}) \quad [12]$$

where $H(f)$ has matrix form $(f_{i+j+1})_{i,j=0}^{\infty}$, and $\tilde{g}(e^{i\theta}) = g(e^{-i\theta})$. Our Banach algebra conditions show that if f is in \mathcal{B} then the operator $H(f)$ satisfies $\sum_{i,j} |a_{ij}|^2 < \infty$, where the a_{ij} are the matrix entries of the operator. Any operator satisfying this is called a Hilbert–Schmidt operator, and it is known that the product of two Hilbert–Schmidt is trace class. Applying the identity to

$$T(\phi_+^{-1})T(\phi_-)$$

shows that this operator is $T(\phi_+^{-1}\phi_-)$ plus trace class. The operator

$$T(\phi_+)T(\phi_-^{-1})$$

is thus $T(\phi_+\phi_-^{-1})$ plus trace class and one more application of the identity combined with the fact that trace class operators form an ideal yield the desired result.

From the theory of infinite determinants, as $n \rightarrow \infty$,

$$\det P_n T(\phi_+^{-1}) T(\phi) T(\phi_-^{-1}) P_n \quad [13]$$

converges to

$$\det(T(\phi_+^{-1}) T(\phi) T(\phi_-^{-1})) \quad [14]$$

At this point, we have proved that

$$\begin{aligned} \lim_{n \rightarrow \infty} D_n(\phi) / ((\phi_-)_0)^n ((\phi_+)_0)^n \\ = \lim_{n \rightarrow \infty} \det P_n T(\phi_+^{-1}) T(\phi) T(\phi_-^{-1}) P_n \\ = \det(T(\phi_+^{-1}) T(\phi) T(\phi_-^{-1})) \end{aligned} \quad [15]$$

It only remains to identify the constants. To see that

$$G(\phi) = ((\phi_-)_0)^n ((\phi_+)_0)^n$$

we note that

$$\begin{aligned} G(\phi) &= \exp((\log \phi)_0) = \exp\left(\frac{1}{2\pi} \int_0^{2\pi} \log \phi(e^{i\theta}) d\theta\right) \\ &= \exp\left(\frac{1}{2\pi} \int_0^{2\pi} (\log \phi_-(e^{i\theta}) + \log \phi(e^{i\theta})) d\theta\right) \\ &= \exp(\log \phi_-)_0 \exp(\log \phi_+)_0 = (\phi_-)_0 (\phi_+)_0 \end{aligned}$$

To compute the determinant of

$$T(\phi_+^{-1}) T(\phi) T(\phi_-^{-1})$$

we write

$$\begin{aligned} \det T(\phi_+^{-1}) T(\phi) T(\phi_-^{-1}) \\ = \det T(\phi_+^{-1}) T(\phi_- \phi_+) T(\phi_-^{-1}) \\ = \det T(\phi_+^{-1}) T(\phi_-) T(\phi_+) T(\phi_-^{-1}) \end{aligned}$$

This last expression is the form

$$e^A e^B e^{-A} e^{-B}$$

where

$$A = -T(\log \phi_+) \quad \text{and} \quad B = T(\log \phi_-)$$

If $AB - BA$ is trace class then

$$\det e^A e^B e^{-A} e^{-B} = e^{\text{tr}(AB - BA)}$$

The operator $AB - BA$ is

$$-T(\log \phi_+) T(\log \phi_-) + T(\log \phi_-) T(\log \phi_+)$$

which equals

$$-T(\log \phi_+) T(\log \phi_-) + T((\log \phi_-)(\log \phi_+))$$

and, by the identity from eqn [12], becomes

$$H(\log \phi_-) H(\log \phi_+)$$

It can be directly computed that

$$\text{tr}(H(\log \phi_-) H(\log \phi_+))$$

equals

$$\sum_{k=1}^{\infty} k s_k s_{-k}$$

and the theorem is proved.

Returning to the Ising model, one needs to compute the asymptotics of the determinants for the generating function

$$\phi(e^{i\theta}) = \left(\frac{(1 - \alpha_1 e^{i\theta})(1 - \alpha_2 e^{-i\theta})}{(1 - \alpha_1 e^{-i\theta})(1 - \alpha_2 e^{i\theta})} \right)^{1/2}$$

The term $G(\phi) = 1$ and for $k > 0$

$$k s_k s_{-k} = \frac{1}{4} \left(\frac{-\alpha_1^{2k}}{k} + \frac{-\alpha_2^{2k}}{k} + 2 \frac{-\alpha_1^k \alpha_2^k}{k} \right)$$

from which it follows that

$$\lim_{n \rightarrow \infty} D_n(\phi) = \left[\frac{(1 - \alpha_1^2)(1 - \alpha_2^2)}{(1 - \alpha_1 \alpha_2)^2} \right]^{1/4}$$

Recalling the definition of α_1 and α_2 yields

$$\lim_{n \rightarrow \infty} \langle \sigma_{0,0} \sigma_{0,n} \rangle = \left[1 - \frac{1}{(\sinh 2\beta E_1 \sinh 2\beta E_2)^2} \right]^{1/4}$$

or the spontaneous magnetization M as

$$M = \left(1 - \frac{1}{(\sinh 2\beta E_1 \sinh 2\beta E_2)^2} \right)^{1/8}$$

In order for this computation to be valid, it was necessary for $0 < \alpha_2 < 1$, and by elementary computations one can show that this is equivalent to the inequality

$$\sinh 2\beta E_1 \sinh 2\beta E_2 > 1$$

Nonsmooth Symbols or $T = T_c$

A problem occurs in the analysis just outlined when the inequality $0 < \alpha_2 < 1$ does not hold. There are two separate possibilities, $\alpha_2 > 1$ or $\alpha_2 = 1$. First, we consider the latter case. For fixed E_1 and E_2 , this happens for exactly one fixed value of the constant $\beta_c = 1/kT_c$ and the corresponding temperature T_c is called the critical temperature. The “strong Szegő

limit theorem” does not apply since our generating function is of the form

$$\phi(e^{i\theta}) = \left(\frac{(1 - \alpha_1 e^{i\theta})(1 - e^{-i\theta})}{(1 - \alpha_1 e^{-i\theta})(1 - e^{i\theta})} \right)^{1/2} \tag{16}$$

In 1968, Fisher and Hartwig raised a conjecture about $D_n(\phi)$ for nonsmooth ϕ which included the above example. They considered generating functions of the form

$$\phi(e^{i\theta}) = \psi(e^{i\theta}) \prod_{j=1}^R \phi_{\alpha_j, \beta_j}(e^{i(\theta - \theta_j)}) \tag{17}$$

where

$$\phi_{\alpha, \beta}(e^{i\theta}) = (2 - 2 \cos \theta)^\alpha e^{i\beta(\theta - \pi)}, \quad 0 < \theta < 2\pi$$

$\Re \alpha > -1/2$, and β is not an integer. The function ψ is assumed to be a smooth function. Using the Fisher–Hartwig notation, the symbol of interest in the Ising model from eqn [16] can be written as

$$\psi(e^{i\theta}) \phi_{0, -1/2}(e^{i\theta})$$

where

$$\psi(e^{i\theta}) = \left(\frac{1 - \alpha_1 e^{i\theta}}{1 - \alpha_1 e^{-i\theta}} \right)^{1/2}$$

The conjecture of Fisher and Hartwig for general symbols of this type stated that

$$D_n(\phi) \sim G(\psi)^n n^p E$$

where

$$p = \sum_{r=1}^R (\alpha_r^2 - \beta_r^2)$$

and E^* is a constant whose value they did not identify. The constant was later computed to be

$$\begin{aligned} E^*(\phi) &= E(\psi) \prod_{j=1}^R \psi_+(e^{i\theta_j})^{-\alpha_j + \beta_j} \psi_-(e^{i\theta_j})^{-\alpha_j - \beta_j} \\ &\times \prod_{1 \leq s \neq r \leq R} (1 - e^{i(\theta_s - \theta_r)})^{-(\alpha_s + \beta_s)(\alpha_r - \beta_r)} \\ &\times \prod_{j=1}^R \frac{G(1 + \alpha_j + \beta_j) G(1 + \alpha_j - \beta_j)}{G(1 + 2\alpha_j)} \end{aligned}$$

where $G(z)$ is the Barnes G -function satisfying

$$G(1 + z) = \Gamma(z) G(z)$$

and is defined by

$$\begin{aligned} G(1 + z) &= (2\pi)^{z/2} e^{-(z+1)z/2 - \gamma z^2/2} \\ &\times \prod_{k=1}^\infty \left(1 + \frac{z}{k} \right)^k e^{-z + z^2/2k} \end{aligned}$$

For the above factors, we normalize ψ so that the geometric mean is 1. Then we may assume that the factors $\psi_+, \psi_- (\psi_+ \psi_- = \psi)$ are 1 at zero and infinity, respectively, and this defines the logarithms for the first product. The $E(\psi)$ term is the constant in Szegő’s theorem, and the argument of a term of the form $(1 - e^{i(\theta_s - \theta_r)})$ is taken between $-\pi/2$ and $\pi/2$.

In the case where $R = 1$, the conjecture is known to hold if $\Re \alpha > -1/2$ and the function b satisfies the conditions of Szegő’s theorem and is infinitely differentiable. The theorem also has an extension to the case where $\Re \alpha < -1/2$, with 2α not an integer, as long as the Fourier coefficients are defined as the coefficients of a distribution.

If we apply the theorem to the generating function from [16]

$$\psi(e^{i\theta}) \phi_{0, -1/2}(e^{i\theta}) = \left(\frac{1 - \alpha_1 e^{i\theta}}{1 - \alpha_1 e^{-i\theta}} \right)^{1/2} \phi_{0, -1/2}(e^{i\theta})$$

we see that the asymptotic expansion is given by

$$n^{-1/4} \left(\frac{1 + \alpha_1}{1 - \alpha_1} \right)^{1/4} G(1/2) G(3/2)$$

This last formula shows that, at the critical temperature,

$$\lim_{n \rightarrow \infty} \langle \sigma_{0,0} \sigma_{0,n} \rangle = \lim_{n \rightarrow \infty} D_n(\phi) = 0$$

thus, $M = 0$, and hence there is no correlation between distant lattice points.

It should be remarked here that the diagonal correlation at the critical temperature is also given by a singular Toeplitz determinant,

$$\langle \sigma_{0,0} \sigma_{n,n} \rangle = D_n(\phi_{0, -1/2}) \sim n^{-1/4} G(1/2) G(3/2)$$

and thus this limit is also zero.

The proof of the Fisher–Hartwig conjecture is much more complicated than the proof of the “strong Szegő limit theorem.” For an indication of how it is proved, note that if we consider the generating function $\phi_{0, \beta}$, the Fourier coefficients are $(\sin \pi \beta) / [\pi(n - \beta)]$ and hence the matrix is Cauchy and the determinant can be computed exactly. From this the asymptotics can be derived and they yield a special case of the Fisher–Hartwig conjecture. The main idea in extending the result to a symbol of the form

$$\psi(e^{i\theta}) \phi_{0, \beta}(e^{i\theta})$$

is to prove that the limit of

$$\frac{D_n(\psi \phi_{0, \beta})}{D_n(\psi) D_n(\phi_{0, \beta})}$$

exists. The proof uses much of the same trace-class approach used in proving the “strong Szegő limit theorem,” although the results are more complicated. These ideas are then extended for $R > 1$ and also more general β and α .

It should be noted that in this article the Fisher–Hartwig conjecture does not always hold. If we consider the function

$$\phi(e^{i\theta}) = \begin{cases} -1, & -\pi < \theta < 0 \\ 1, & 0 < \theta < \pi \end{cases}$$

then

$$\phi_n = \begin{cases} 0, & \text{if } n \text{ is even} \\ -2i/(\pi k), & \text{if } n \text{ is odd} \end{cases}$$

The matrix $T_n(\phi)$ is antisymmetric and, if n is odd, $D_n(\phi) = 0$. If n is even, using elementary row and column operations, the determinant can be put in block form with each block of Cauchy type. The determinant can then be evaluated to find

$$D_n(\phi) \sim (i)^n n^{-1/2} K$$

where K is a certain constant.

It is instructive to note that

$$\begin{aligned} \phi(e^{i\theta}) &= \phi_{0,1/2}(e^{i\theta})\phi_{0,-1/2}(e^{i(\theta-\pi)}) \\ &= \phi_{0,-1/2}(e^{i\theta})\phi_{0,1/2}(e^{i(\theta-\pi)}) \end{aligned}$$

and thus that this particular symbol has two representations of the type given in [17] and each would give a different asymptotic expansion of the determinant if the conjecture were true for this set of parameters. Hence, it is clear that the conjecture must fail to hold in this case.

However, this example indicates that there might be a generalization of the original conjecture of Fisher and Hartwig. If

$$\phi_{\alpha,\beta}(e^{i(\theta-\gamma)}) = \phi_{\alpha,\beta,\gamma}$$

then

$$\phi = \psi \prod_{j=1}^R \phi_{\alpha_j, \beta_j, \theta_j}$$

it is also the case that

$$\phi = \psi^* \prod_{j=1}^R \phi_{\alpha_j, \beta_j + n_j, \theta_j}$$

where

$$\sum_{j=1}^R n_j = 0 \quad \text{and} \quad \psi^* = \psi \prod_{j=1}^R (-e^{i\theta_j})^{n_j}$$

In the example above, $\beta_1 = 1/2$, $\beta_2 = -1/2$, $\theta_1 = 0$, $\theta_2 = \pi$, $n_1 = -1$, and $n_2 = 1$. The result for the counterexample, combined with what is known for the case of integer values of α and β , leads to the following generalized conjecture. Suppose

$$\psi(e^{i\theta}) = \phi^k \prod_{j=1}^R \phi_{\alpha_j^k, \beta_j^k, \theta_j}$$

for some set of indices k . Define $Q(k) = \sum_{j=1}^R (\alpha_j^k)^2 - (\beta_j^k)^2$. Let $Q = \max_k \Re(Q(k))$ and

$$\mathcal{K} = \{k \mid \Re(Q(k)) = Q\}$$

The generalized asymptotic formula is conjectured to be

$$D_n(\psi) = \sum_{k \in \mathcal{K}} G(\psi^k) n^{Q(k)} E_k^* + o(|G(\phi)|^n n^Q)$$

It may turn out that there is only one element in \mathcal{K} and for these symbols there is a unique representation that yields the highest power in the exponent of the asymptotic expansion. These are the symbols for which the original Fisher–Hartwig conjecture should be true and it is now confirmed in these cases. For example, the conjecture is known to hold for $R > 1$ when $|\Re \alpha_r| < 1/2$ and $|\Re \beta_r| < 1/2$.

Symbols with Nonzero Index or $T > T_c$

The last possibility in computing the correlation asymptotics is the case where $\alpha_2 > 1$. Note that, for fixed E_1 and E_2 , there is exactly one value of $\beta = 1/kT$ where

$$\alpha_2 = z_1^{-1} \left(\frac{1 - z_2}{1 + z_2} \right) = 1$$

For values of $T > T_c$, we have that the symbol

$$\left(\frac{(1 - \alpha_1 e^{i\theta})(1 - \alpha_2 e^{-i\theta})}{(1 - \alpha_1 e^{-i\theta})(1 - \alpha_2 e^{i\theta})} \right)^{1/2}$$

is the same as

$$e^{-i\theta} \left(\frac{(1 - \alpha_1 e^{i\theta})(1 - (1/\alpha_2) e^{i\theta})}{(1 - \alpha_1 e^{-i\theta})(1 - (1/\alpha_2) e^{-i\theta})} \right)^{1/2}$$

with the argument chosen so that the symbol is positive at π . Except for the extra factor of $e^{-i\theta}$, this is the same type of smooth symbol that was considered earlier (see the section “Strong Szegő limit theorem”). However, a factor of $e^{i\theta}$ can change the asymptotics considerably as can be seen by considering the simple example of the $\phi \equiv 1$. Fortunately, a variation of the Szegő theorem, first

considered by Fisher and Hartwig, holds for this case of smooth, nonvanishing index.

Theorem 2 Suppose that $\phi = \phi_- \phi_+$ satisfies the condition of the “strong Szegő limit theorem” and in addition is at least once continuously differentiable. Then, if $b = \phi_- \phi_+^{-1}$ and $c = \phi_-^{-1} \phi_+$,

$$D_n(e^{im\theta} \phi) \sim (-1)^{(n-1+m)m} G(\phi)^n E(\phi) G(c)^m \tag{18}$$

$$\begin{aligned} &\times \left(\det \begin{pmatrix} b_n & \cdots & b_{n-m+1} \\ \vdots & \ddots & \vdots \\ b_{n-1+m} & \cdots & b_n \end{pmatrix} + O(n^{-3}) \right) \\ &\times (1 + O(n^{-1})) \end{aligned} \tag{19}$$

Applying this to the symbol

$$e^{-i\theta} \phi(e^{i\theta}) = e^{-i\theta} \left(\frac{(1 - \alpha_1 e^{i\theta})(1 - (1/\alpha_2) e^{i\theta})}{(1 - \alpha_1 e^{-i\theta})(1 - (1/\alpha_2) e^{-i\theta})} \right)^{1/2}$$

we have that $m = 1$, $G(\phi) = -1$, $G(c) = -1$, and

$$E(\phi) = \left[(1 - \alpha_1^2) \left(1 - \frac{1}{\alpha_2^2} \right) \left(1 - \frac{\alpha_1}{\alpha_2} \right)^2 \right]^{1/4} \tag{20}$$

The determinant in the above formula is the constant

$$\begin{aligned} b_n &= \frac{1}{2\pi} \int_0^{2\pi} \left((1 - \alpha_1 e^{i\theta}) \left(1 - \frac{1}{\alpha_2} e^{i\theta} \right) \right. \\ &\quad \left. \times (1 - \alpha_1 e^{-i\theta}) \left(1 - \frac{1}{\alpha_2} e^{-i\theta} \right) \right)^{-1/2} e^{-in\theta} d\theta \end{aligned}$$

The last integral can be deformed to a segment of the real line and evaluated asymptotically to find that the leading term is

$$\begin{aligned} &-\frac{1}{\sqrt{\pi} \alpha_2^n} \left(\left(1 - \frac{\alpha_1}{\alpha_2} \right) (1 - \alpha_1 \alpha_2) \left(1 - \frac{1}{\alpha_2^2} \right) \right)^{-1/2} \\ &\times \frac{\Gamma(n + 1/2)}{\Gamma(n + 1)} \end{aligned}$$

Putting this together with the above constants, we have, for $T > T_c$,

$$\begin{aligned} &\langle \sigma_{0,0} \sigma_{0,n} \rangle \\ &\sim \frac{1}{\sqrt{\pi n} \alpha_2^n} \left[(1 - \alpha_1^2)^{1/4} \left(1 - \frac{1}{\alpha_2^2} \right)^{-1/4} (1 - \alpha_1 \alpha_2)^{-1/2} \right] \end{aligned}$$

This implies that the correlation tends to zero very rapidly as $n \rightarrow \infty$.

Further Remarks

The interaction between statistical mechanics and the theory of Toeplitz determinants has a long history, and much of the motivation to describe the asymptotics of the determinants was spurred by the question of spontaneous magnetization in the two-dimensional Ising model. The previous three sections attempt to show how the very different physical situations – $T < T_c$, $T = T_c$, and $T > T_c$ – all correspond to very different behavior in the symbols of the generating functions. Critical systems predict qualitatively different Szegő type theorems. For example, the phase transition at T_c predicts that the asymptotics for singular symbols cannot be predicted by the smooth symbols, that is, one cannot use continuous functions to approximate the results for singular symbols.

Onsager (1971) was the first to understand that the correlation function could be expressed as a Toeplitz determinant. This was made explicit by Montroll *et al.* (1963). For more information about the Ising model, the reader is referred to McCoy and Wu (1973), where a clear and complete description of the Ising model (and most of the notation used here in reference to this model) can be found.

Szegő (1915, 1952) had originally proved a weak form of the “limit” theorem and he understood that it was desirable to extend to a second-order term. Szegő first proved the “strong Szegő limit theorem” for positive generating functions and this was later extended to the nonpositive case.

The first to understand that a different asymptotic behavior was expected at the critical temperature was Fisher and this resulted in the conjecture for the class of determinants generated by what is now known as Fisher–Hartwig symbols (Fisher and Hartwig 1968). Progress on the conjecture was made by many authors. Böttcher and Silbermann (1998) have provided general results concerning Toeplitz operators and determinants. Additional information about the conjectures of Fisher and Hartwig can be found in Böttcher and Silbermann (1990, 1998), Ehrhardt (2001), and Ehrhardt and Silbermann (1997).

Toeplitz determinants are also important in many other applications. One more recent area of interest is the connection between random-matrix theory and Toeplitz determinants. Many statistical quantities for the circular unitary ensemble can be described as a Toeplitz determinant. For example, the probability of finding no eigenvalues in an interval can be expressed as a Toeplitz determinant. It is also the case that many of the most interesting

statistics correspond to singular symbols. For basic random-matrix theory information see Mehta (1991), and for connections between the circular unitary ensemble and Toeplitz determinants, see Hughes (2001), Tracy and Widom (1993), and Widom (1994).

See also: Integrable Systems in Random Matrix Theory; Two-Dimensional Ising Model.

Further Reading

- Böttcher A and Silbermann B (1990) *Analysis of Toeplitz Operators*. Berlin: Akademie-Verlag.
- Böttcher A and Silbermann B (1998) *Introduction to Large Truncated Toeplitz Matrices*. Berlin: Springer.
- Ehrhardt T (2001) A status report on the asymptotic behavior of Toeplitz determinants with Fisher–Hartwig singularities. *Operator Theory: Advances and Applications* 124: 217–241.
- Ehrhardt T and Silbermann B (1997) Toeplitz determinants with one Fisher–Hartwig singularity. *Journal of Functional Analysis* 148: 229–256.
- Fisher ME and Hartwig RE (1968) Toeplitz determinants: some applications, theorems, and conjectures. *Advances in Chemical Physics* 15: 333–353.

- Hughes C, Keating JP, and O’Connell N (2001) On the characteristic polynomial of a random unitary matrix. *Communications in Mathematical Physics* 220: 429–451.
- McCoy BM and Wu TT (1973) *The Two-Dimensional Ising Model*. Cambridge, MA: Harvard University Press.
- Metha ML (1991) *Random Matrices*. San Diego: Academic Press.
- Montroll EW, Potts RB, and Ward JC (1963) Correlations and spontaneous magnetization of the two-dimensional Ising model. *Journal of Mathematical Physics* 4(2): 308–322.
- Onsager L (1971) The Ising model in two dimensions. In: Mills RE, Ascher E, and Jaffee RI (eds.) *Critical Phenomena in Alloys, Magnets, and Superconductors*, pp. 3–12. New York: McGraw-Hill.
- Szegő G (1915) Ein Grenzwertsatz über die Toeplitzschen Determinanten einer reellen positiven Funktion. *Mathematische Annalen* 76: 490–503.
- Szegő G (1952) On Certain Hermitian Forms Associated with the Fourier Series of a Positive Function, pp. 222–238. Lund: Festschrift Marcel Riesz.
- Tracy CA and Widom H (1993) Introduction to random matrices. In: Helminck GF (ed.) *Geometric and Quantum Aspects of Integrable Systems*, (Scheveningen, 1992), Lecture Notes in Physics, vol. 424, pp. 103–130. Berlin: Springer.
- Widom H (1994) Random Hermitian matrices and (nonrandom) Toeplitz matrices. In: Basor E and Gohberg I (eds.) *Toeplitz Operators and Related Topics* (Santa Cruz, CA, 1992), Oper. Theory Adv. Appl., vol. 71, pp. 9–15. Basel: Birkhäuser.

Tomita–Takesaki Modular Theory

S J Summers, University of Florida, Gainesville, FL, USA

© 2006 Elsevier Ltd. All rights reserved.

Basic Structure

The origins of Tomita–Takesaki modular theory lie in two unpublished papers of M Tomita in 1967 and a slim volume by Takesaki (1970). It has developed into one of the most important tools in the theory of operator algebras and has found many applications in mathematical physics.

Although the modular theory has been formulated in a more general setting, it will be presented in the form in which it most often finds application in mathematical physics (for generalizations, details, and further references concerning the material covered in this article, the reader is referred to the Further Reading section). Let \mathcal{M} be a von Neumann algebra on a Hilbert space \mathcal{H} containing a vector Ω which is cyclic and separating for \mathcal{M} . Define the operator S_0 on \mathcal{H} as follows:

$$S_0 A \Omega = A^* \Omega, \quad \text{for all } A \in \mathcal{M}$$

This operator extends to a closed antilinear operator S defined on a dense subset of \mathcal{H} . Let Δ be the unique positive, self-adjoint operator and J the

unique antiunitary operator occurring in the polar decomposition

$$S = J \Delta^{1/2} = \Delta^{-1/2} J$$

Δ is called the modular operator and J the modular conjugation (or modular involution) associated with the pair (\mathcal{M}, Ω) . Note that J^2 is the identity operator and $J = J^*$. Moreover, the spectral calculus may be applied to Δ so that Δ^{it} is a unitary operator for each $t \in \mathbb{R}$ and $\{\Delta^{it} \mid t \in \mathbb{R}\}$ forms a strongly continuous unitary group. Let \mathcal{M}' denote the set of all bounded linear operators on \mathcal{H} which commute with all elements of \mathcal{M} . The modular theory begins with the following remarkable theorem.

Theorem 1 *Let \mathcal{M} be a von Neumann algebra with a cyclic and separating vector Ω . Then $J\Omega = \Omega = \Delta\Omega$, and the following equalities hold:*

$$J\mathcal{M}J = \mathcal{M}'$$

and

$$\Delta^{it} \mathcal{M} \Delta^{-it} = \mathcal{M}, \quad \text{for all } t \in \mathbb{R}$$

Note that if one defines $F_0 A' \Omega = A'^* \Omega$, for all $A' \in \mathcal{M}'$, and takes its closure F , then one has the relations

$$\Delta = FS, \quad \Delta^{-1} = SF, \quad F = J \Delta^{-1/2}$$

Modular Automorphism Group

By Theorem 1, the unitaries $\Delta^{it}, t \in \mathbb{R}$, induce a one-parameter automorphism group $\{\sigma_t\}$ of \mathcal{M} by

$$\sigma_t(A) = \Delta^{it} A \Delta^{-it}, \quad A \in \mathcal{M}, \quad t \in \mathbb{R}$$

This group is called the modular automorphism group of \mathcal{M} (relative to Ω). Let ω denote the faithful normal state on \mathcal{M} induced by Ω :

$$\omega(A) = \frac{1}{\|\Omega\|^2} \langle \Omega, A\Omega \rangle, \quad A \in \mathcal{M}$$

From Theorem 1 it follows that ω is invariant under $\{\sigma_t\}$, that is, $\omega(\sigma_t(A)) = \omega(A)$ for all $A \in \mathcal{M}$ and $t \in \mathbb{R}$.

The modular automorphism group contains information about both \mathcal{M} and ω . For example, the modular automorphism group is an inner automorphism on \mathcal{M} if and only if \mathcal{M} is semifinite. It is trivial if and only if ω is a tracial state on \mathcal{M} . Indeed, for any $B \in \mathcal{M}$, one has $\sigma_t(B) = B$ for all $t \in \mathbb{R}$ if and only if $\omega(AB) = \omega(BA)$ for all $A \in \mathcal{M}$. Let \mathcal{M}^σ denote the set of all such B in \mathcal{M} .

The KMS Condition

The modular automorphism group satisfies a condition which had already been used in mathematical physics to characterize equilibrium temperature states of quantum systems in statistical mechanics and field theory – the Kubo–Martin–Schwinger (KMS) condition. If \mathcal{M} is a von Neumann algebra and $\{\alpha_t | t \in \mathbb{R}\}$ is a σ -weakly continuous one-parameter group of automorphisms of \mathcal{M} , then the state ϕ on \mathcal{M} satisfies the KMS condition at (inverse temperature) β ($0 < \beta < \infty$) with respect to $\{\alpha_t\}$ if for any $A, B \in \mathcal{M}$ there exists a complex function $F_{AB}(z)$ which is analytic on the strip $\{z \in \mathbb{C} | 0 < \text{Im } z < \beta\}$ and continuous on the closure of this strip such that

$$\begin{aligned} F_{A,B}(t) &= \phi(\alpha_t(A)B) \\ F_{A,B}(t + i\beta) &= \phi(B\alpha_t(A)) \end{aligned}$$

for all $t \in \mathbb{R}$. In this case, $\phi(\alpha_{i\beta}(A)B) = \phi(BA)$, for all A, B in a σ -weakly dense, α -invariant $*$ -subalgebra of \mathcal{M} . Such KMS states are α -invariant, that is, $\phi(\alpha_t(A)) = \phi(A)$, for all $A \in \mathcal{M}, t \in \mathbb{R}$, and are stable and passive (cf. Bratteli and Robinson (1981) and Haag (1992)).

Every faithful normal state satisfies the KMS condition at $\beta = 1$ (henceforth called the modular condition) with respect to the corresponding modular automorphism group.

Theorem 2 *Let \mathcal{M} be a von Neumann algebra with a cyclic and separating vector Ω . Then the induced state ω on \mathcal{M} satisfies the modular condition with respect to the modular automorphism group $\{\sigma_t | t \in \mathbb{R}\}$ associated to the pair (\mathcal{M}, Ω) .*

The modular automorphism group is, therefore, endowed with the analyticity associated with the KMS condition, and this is a powerful tool in many applications of the modular theory to mathematical physics. In addition, the physical properties and interpretations of KMS states are often invoked when applying modular theory to quantum physics.

Note that while the nontriviality of the modular automorphism group gives a measure of the non-tracial nature of the state, the KMS condition for the modular automorphism group provides the missing link between the values $\omega(AB)$ and $\omega(BA)$, for all $A, B \in \mathcal{M}$ (hence the use of the term “modular,” as in the theory of integration on locally compact groups).

The modular condition is quite restrictive. Only the modular group can satisfy the modular condition for (\mathcal{M}, Ω) , and the modular group for one state can satisfy the modular condition only in states differing from the original state by the action of an element in the center of \mathcal{M} .

Theorem 3 *Let \mathcal{M} be a von Neumann algebra with a cyclic and separating vector Ω , and let $\{\sigma_t\}$ be the corresponding modular automorphism group. If the induced state ω satisfies the modular condition with respect to a group $\{\alpha_t\}$ of automorphisms of \mathcal{M} , then $\{\alpha_t\}$ must coincide with $\{\sigma_t\}$. Moreover, a normal state ψ on \mathcal{M} satisfies the modular condition with respect to $\{\sigma_t\}$ if and only if $\psi(\cdot) = \omega(h \cdot) = \omega(h^{1/2} \cdot h^{1/2})$ for some unique positive injective operator h affiliated with the center of \mathcal{M} .*

Hence, if \mathcal{M} is a factor, two distinct states cannot share the same modular automorphism group. The relation between the modular automorphism groups for two different states will be described in more detail.

One Algebra and Two States

Consider a von Neumann algebra \mathcal{M} with two cyclic and separating vectors Ω and Φ , and denote by ω and ϕ , respectively, the induced states on \mathcal{M} . Let $\{\sigma_t^\omega\}$ and $\{\sigma_t^\phi\}$ denote the corresponding modular groups. There is a general relation between the modular automorphism groups of these states.

Theorem 4 *There exists a σ -strongly continuous map $\mathbb{R} \ni t \mapsto U_t \in \mathcal{M}$ such that*

- (i) U_t is unitary for all $t \in \mathbb{R}$;
- (ii) $U_{t+s} = U_t \sigma_t^\omega(U_s)$ for all $s, t \in \mathbb{R}$; and
- (iii) $\sigma_t^\phi(A) = U_t \sigma_t^\omega(A) U_t^*$ for all $A \in \mathcal{M}$ and $t \in \mathbb{R}$.

The 1-cocycle $\{U_t\}$ is commonly called the cocycle derivative of ϕ with respect to ω and one writes $U_t = (D\phi : D\omega)_t$. There is a chain rule for this derivative, as well: If ϕ, ψ , and ρ are faithful normal states on \mathcal{M} , then $(D\psi : D\phi)_t = (D\psi : D\rho)_t (D\rho : D\phi)_t$, for all $t \in \mathbb{R}$. More can be said about the cocycle derivative if the states satisfy any of the conditions in the following theorem.

Theorem 5 *The following conditions are equivalent:*

- (i) ϕ is $\{\sigma_t^\omega\}$ -invariant;
- (ii) ω is $\{\sigma_t^\phi\}$ -invariant;
- (iii) *there exists a unique positive injective operator h affiliated with $\mathcal{M}^{\sigma^\omega} \cap \mathcal{M}^{\sigma^\phi}$ such that $\omega(\cdot) = \phi(h \cdot) = \phi(h^{1/2} \cdot h^{1/2})$;*
- (iv) *there exists a unique positive injective operator h' affiliated with $\mathcal{M}^{\sigma^\omega} \cap \mathcal{M}^{\sigma^\phi}$ such that $\phi(\cdot) = \omega(h' \cdot) = \omega(h'^{1/2} \cdot h'^{1/2})$;*
- (v) *the norms of the linear functionals $\omega + i\phi$ and $\omega - i\phi$ are equal; and*
- (vi) $\sigma_t^\omega \sigma_s^\phi = \sigma_s^\phi \sigma_t^\omega$, for all $s, t \in \mathbb{R}$.

The conditions in Theorem 5 turn out to be equivalent to the cocycle derivative being a representation.

Theorem 6 *The cocycle $\{U_t\}$ intertwining $\{\sigma_t^\omega\}$ with $\{\sigma_t^\phi\}$ is a group representation of the additive group of reals if and only if ϕ and ω satisfy the conditions in Theorem 5. In that case, $U(t) = h^{-it}$.*

The operator $h' = h^{-1}$ in Theorem 5 is called the Radon–Nikodym derivative of ϕ with respect to ω (often denoted by $d\phi/d\omega$), due to the following result, which, if the algebra \mathcal{M} is abelian, is the well-known Radon–Nikodym theorem from measure theory.

Theorem 7 *If ϕ and ω are normal positive linear functionals on \mathcal{M} such that $\phi(A) \leq \omega(A)$, for all positive elements $A \in \mathcal{M}$, then there exists a unique element $h^{1/2} \in \mathcal{M}$ such that $\phi(\cdot) = \omega(h^{1/2} \cdot h^{1/2})$ and $0 \leq h^{1/2} \leq 1$.*

The analogies with measure theory are not accidental, although these are not discussed in detail here. Indeed, any normal trace on a (finite) von Neumann algebra \mathcal{M} gives rise to a noncommutative integration theory in a natural manner. Modular theory affords an extension of this theory to the

setting of faithful normal functionals η on von Neumann algebras \mathcal{M} of any type, enabling the definition of noncommutative L^p spaces, $L^p(\mathcal{M}, \eta)$.

Modular Invariants and the Classification of von Neumann Algebras

As already mentioned, the modular structure carries information about the algebra. This is best evidenced in the structure of type III factors. As this theory is rather involved, only a sketch of some of the results can be given.

If \mathcal{M} is a type III algebra, then its crossed product $\mathcal{N} = \mathcal{M} \rtimes_{\sigma^\omega} \mathbb{R}$ relative to the modular automorphism group of any faithful normal state ω on \mathcal{M} is a type II $_\infty$ algebra with a faithful semifinite normal trace τ such that $\tau \circ \theta_t = e^{-t} \tau$, $t \in \mathbb{R}$, where θ is the dual of σ^ω on \mathcal{N} . Moreover, the algebra \mathcal{M} is isomorphic to the cross product $\mathcal{N} \rtimes_{\theta} \mathbb{R}$, and this decomposition is unique in a very strong sense. This structure theorem entails the existence of important algebraic invariants for \mathcal{M} , which has many consequences, one of which is made explicit here.

If ω is a faithful normal state of a von Neumann algebra \mathcal{M} induced by Ω , let Δ_ω denote the modular operator associated to (\mathcal{M}, Ω) and $\text{sp } \Delta_\omega$ denote the spectrum of Δ_ω . The intersection

$$S'(\mathcal{M}) = \cap \text{sp } \Delta_\omega$$

over all faithful normal states ω of \mathcal{M} is an algebraic invariant of \mathcal{M} .

Theorem 8 *Let \mathcal{M} be a factor acting on a separable Hilbert space. If \mathcal{M} is of type III, then $0 \in S'(\mathcal{M})$; otherwise, $S'(\mathcal{M}) = \{0, 1\}$ if \mathcal{M} is of type I $_\infty$ or II $_\infty$ and $S'(\mathcal{M}) = \{1\}$ if not. Let \mathcal{M} now be a factor of type III.*

- (i) \mathcal{M} is of type III $_\lambda$, $0 < \lambda < 1$, if and only if $S'(\mathcal{M}) = \{0\} \cup \{\lambda^n \mid n \in \mathbb{Z}\}$.
- (ii) \mathcal{M} is of type III $_0$ if and only if $S'(\mathcal{M}) = \{0, 1\}$.
- (iii) \mathcal{M} is of type III $_1$ if and only if $S'(\mathcal{M}) = [0, \infty)$.

In certain physically relevant situations, the spectra of the modular operators of all faithful normal states coincide, so that Theorem 8 entails that it suffices to compute the spectrum of any conveniently chosen modular operator in order to determine the type of \mathcal{M} . In other such situations, there are distinguished states ω such that $S'(\mathcal{M}) = \text{sp } \Delta_\omega$. One such example is provided by asymptotically abelian systems. A von Neumann algebra \mathcal{M} is said to be “asymptotically abelian” if there exists a sequence $\{\alpha_n\}_{n \in \mathbb{N}}$ of automorphisms of \mathcal{M} such that the limit of $\{A\alpha_n(B) - \alpha_n(B)A\}_{n \in \mathbb{N}}$ in

the strong operator topology is zero, for all $A, B \in \mathcal{M}$. If the state ω is α_n -invariant, for all $n \in \mathbb{N}$, then $\text{sp } \Delta_\omega$ is contained in $\text{sp } \Delta_\phi$, for all faithful normal states ϕ on \mathcal{M} , so that $S'(\mathcal{M}) = \text{sp } \Delta_\omega$. If, moreover, $\text{sp } \Delta_\omega = [0, \infty)$, then $\text{sp } \Delta_\omega = \text{sp } \Delta_\phi$, for all ϕ as described.

Self-Dual Cones

Let $j: \mathcal{M} \rightarrow \mathcal{M}'$ denote the antilinear $*$ -isomorphism defined by $j(A) = JAJ, A \in \mathcal{M}$. The natural positive cone \mathcal{P}^\natural associated with the pair (\mathcal{M}, Ω) is defined as the closure, in \mathcal{H} , of the set of vectors

$$\{Aj(A)\Omega \mid A \in \mathcal{M}\}$$

Let \mathcal{M}_+ denote the set of all positive elements of \mathcal{M} . The following theorem collects the main attributes of the natural cone.

Theorem 9

- (i) \mathcal{P}^\natural coincides with the closure in \mathcal{H} of the set $\{\Delta^{1/4}A\Omega \mid A \in \mathcal{M}_+\}$.
- (ii) $\Delta^{it}\mathcal{P}^\natural = \mathcal{P}^\natural$ for all $t \in \mathbb{R}$.
- (iii) $J\Phi = \Phi$ for all $\Phi \in \mathcal{P}^\natural$.
- (iv) $Aj(A)\mathcal{P}^\natural \subset \mathcal{P}^\natural$ for all $A \in \mathcal{M}$.
- (v) \mathcal{P}^\natural is a pointed, self-dual cone whose linear span coincides with \mathcal{H} .
- (vi) If $\Phi \in \mathcal{P}^\natural$, then Φ is cyclic for \mathcal{M} if and only if Φ is separating for \mathcal{M} .
- (vii) If $\Phi \in \mathcal{P}^\natural$ is cyclic, and hence separating, for \mathcal{M} , then the modular conjugation and the natural cone associated with the pair (\mathcal{M}, Φ) coincide with J and \mathcal{P}^\natural , respectively.
- (viii) For every normal positive linear functional ϕ on \mathcal{M} , there exists a unique vector $\Phi_\phi \in \mathcal{P}^\natural$ such that $\phi(A) = \langle \Phi_\phi, A\Phi_\phi \rangle$ for all $A \in \mathcal{M}$.

In fact, the algebras \mathcal{M} and \mathcal{M}' are uniquely characterized by the natural cone \mathcal{P}^\natural [4]. In light of (viii), if α is an automorphism of \mathcal{M} , then

$$V(\alpha)\Phi_\phi = \Phi_{\phi \circ \alpha^{-1}}$$

defines an isometric operator on \mathcal{P}^\natural , which by (v) extends to a unitary operator on \mathcal{H} . The map $\alpha \mapsto V(\alpha)$ defines a unitary representation of the group of automorphisms $\text{Aut}(\mathcal{M})$ on \mathcal{M} in such a manner that $V(\alpha)AV(\alpha)^{-1} = \alpha(A)$ for all $A \in \mathcal{M}$ and $\alpha \in \text{Aut}(\mathcal{M})$. Indeed, one has the following:

Theorem 10 Let \mathcal{M} be a von Neumann algebra with a cyclic and separating vector Ω . The group \mathcal{V} of all unitaries V satisfying

$$VMV^* = \mathcal{M}, \quad VJV^* = J, \quad V\mathcal{P}^\natural = \mathcal{P}^\natural$$

is isomorphic to $\text{Aut}(\mathcal{M})$ under the above map $\alpha \mapsto V(\alpha)$, which is called the “standard implementation” of $\text{Aut}(\mathcal{M})$.

Often of particular physical interest are (anti-)automorphisms of \mathcal{M} leaving ω invariant. They can only be implemented by (anti)unitaries which leave the pair (\mathcal{M}, Ω) invariant. In fact, if U is a unitary or antiunitary operator satisfying $U\Omega = \Omega$ and $UMU^* = \mathcal{M}$, then U commutes with both J and Δ .

Two Algebras and One State

Motivated by applications to quantum field theory, the study of the modular structures associated with one state and more than one von Neumann algebra has begun (see Borchers (2000) for references and details). Let $\mathcal{N} \subset \mathcal{M}$ be von Neumann algebras with a common cyclic and separating vector Ω , and $\Delta_{\mathcal{N}}, J_{\mathcal{N}}$ and $\Delta_{\mathcal{M}}, J_{\mathcal{M}}$ denote the corresponding modular objects. The structure $(\mathcal{M}, \mathcal{N}, \Omega)$ is called a \pm -half-sided modular inclusion if $\Delta_{\mathcal{M}}^{it}\mathcal{N}\Delta_{\mathcal{M}}^{-it} \subset \mathcal{N}$, for all $\pm t \geq 0$.

Theorem 11 Let \mathcal{M} be a von Neumann algebra with cyclic and separating vector Ω . The following are equivalent:

- (i) There exists a proper subalgebra $\mathcal{N} \subset \mathcal{M}$ such that $(\mathcal{M}, \mathcal{N}, \Omega)$ is a \mp -half-sided modular inclusion.
- (ii) There exists a unitary group $\{U(t)\}$ with positive generator such that

$$U(t)\mathcal{M}U(t)^{-1} \subset \mathcal{M}, \quad \text{for all } \pm t \geq 0, \\ U(t)\Omega = \Omega, \quad \text{for all } t \in \mathbb{R}$$

Moreover, if these conditions are satisfied, then the following relations must hold:

$$\Delta_{\mathcal{M}}^{it}U(s)\Delta_{\mathcal{M}}^{-it} = \Delta_{\mathcal{N}}^{it}U(s)\Delta_{\mathcal{N}}^{-it} = U(e^{\mp 2\pi t}s)$$

and

$$J_{\mathcal{M}}U(s)J_{\mathcal{M}} = J_{\mathcal{N}}U(s)J_{\mathcal{N}} = U(-s)$$

for all $s, t \in \mathbb{R}$. In addition, $\mathcal{N} = U(\pm 1)\mathcal{M}U(\pm 1)^{-1}$, and if \mathcal{M} is a factor, it must be type III_1 .

The richness of this structure is further suggested by the next theorem.

Theorem 12

- (i) Let $(\mathcal{M}, \mathcal{N}_1, \Omega)$ and $(\mathcal{M}, \mathcal{N}_2, \Omega)$ be $--$ -half-sided, resp. $+-$ -half-sided, modular inclusions satisfying the condition $J_{\mathcal{N}_1}J_{\mathcal{N}_2} = J_{\mathcal{M}}J_{\mathcal{N}_2}J_{\mathcal{N}_1}J_{\mathcal{M}}$. Then the modular unitaries $\Delta_{\mathcal{M}}^{it}, \Delta_{\mathcal{N}_1}^{is}, \Delta_{\mathcal{N}_2}^{iu}, s, t, u \in \mathbb{R}$, generate a faithful continuous unitary representation of the identity component of the

group of isometries of two-dimensional Minkowski space.

- (ii) Let $\mathcal{M}, \mathcal{N}, \mathcal{N} \cap \mathcal{M}$ be von Neumann algebras with a common cyclic and separating vector Ω . If $(\mathcal{M}, \mathcal{M} \cap \mathcal{N}, \Omega)$ and $(\mathcal{N}, \mathcal{M} \cap \mathcal{N}, \Omega)$ are $-$ -half-sided, resp. $+$ -half-sided, modular inclusions such that $J_{\mathcal{N}} \mathcal{M} J_{\mathcal{N}} = \mathcal{M}$, then the modular unitaries $\Delta_{\mathcal{M}}^{it}, \Delta_{\mathcal{N}}^{is}, \Delta_{\mathcal{N} \cap \mathcal{M}}^{iu}, s, t, u \in \mathbb{R}$, generate a faithful continuous unitary representation of $\mathrm{SL}(2, \mathbb{R})/\mathbb{Z}_2$.

This has led to a further useful notion. If $\mathcal{N} \subset \mathcal{M}$ and Ω is cyclic for $\mathcal{N} \cap \mathcal{M}$, then $(\mathcal{M}, \mathcal{N}, \Omega)$ is said to be a “ \pm -modular intersection” if both $(\mathcal{M}, \mathcal{M} \cap \mathcal{N}, \Omega)$ and $(\mathcal{N}, \mathcal{M} \cap \mathcal{N}, \Omega)$ are \pm -half-sided modular inclusions and

$$J_{\mathcal{N}} \left[\lim_{t \rightarrow \mp \infty} \Delta_{\mathcal{N}}^{it} \Delta_{\mathcal{M}}^{-it} \right] J_{\mathcal{N}} = \lim_{t \rightarrow \mp \infty} \Delta_{\mathcal{M}}^{it} \Delta_{\mathcal{N}}^{-it}$$

where the existence of the strong operator limits is assured by the preceding assumptions. An example of the utility of this structure is the following theorem.

Theorem 13 Let $\mathcal{N}, \mathcal{M}, \mathcal{L}$ be von Neumann algebras with a common cyclic and separating vector Ω . If $(\mathcal{M}, \mathcal{N}, \Omega)$ and $(\mathcal{N}', \mathcal{L}, \Omega)$ are $-$ -modular intersections and $(\mathcal{M}, \mathcal{L}, \Omega)$ is a $+$ -modular intersection, then the unitaries $\Delta_{\mathcal{M}}^{it}, \Delta_{\mathcal{N}}^{is}, \Delta_{\mathcal{L}}^{iu}, s, t, u \in \mathbb{R}$, generate a faithful continuous unitary representation of $\mathrm{SO}^+(1, 2)$.

These results and their extensions to larger numbers of algebras were developed for application in algebraic quantum field theory, but one may anticipate that half-sided modular inclusions will find wider use. Modular theory has also been applied fruitfully in the theory of inclusions $\mathcal{N} \subset \mathcal{M}$ of properly infinite algebras with finite or infinite index.

Applications in Quantum Theory

The Tomita-Takesaki theory has found many applications in quantum field theory and quantum statistical mechanics. As mentioned earlier, the modular automorphism group satisfies the KMS condition, a property of physical significance in the quantum theory of many-particle systems, which includes quantum statistical mechanics and quantum field theory. In such settings, for a suitable algebra of observables \mathcal{M} and state ω , an automorphism group $\{\sigma_{\beta t}\}$ representing the time evolution of the system satisfies the modular condition. Hence, on the one hand, $\{\sigma_{\beta t}\}$ is the modular automorphism group of the pair (\mathcal{M}, Ω) , and, on the other, ω is an

equilibrium state at inverse temperature β , with all the consequences which both of these facts have.

But it has become increasingly clear that the modular objects Δ^{it}, J , of certain algebras of observables and states encode additional physical information. In 1975, it was discovered that if one considers the algebras of observables associated with a finite-component quantum field theory satisfying the Wightman axioms, then the modular objects associated with the vacuum state and algebras of observables localized in certain wedge-shaped regions in Minkowski space have geometric content. In fact, the unitary group $\{\Delta^{it}\}$ implements the group of Lorentz boosts leaving the wedge region invariant (this property is now called modular covariance), and the modular involution J implements the space-time reflection about the edge of the wedge, along with a charge conjugation. This discovery caused some intense research activity (see Baumgartel and Wollenberg 1992, Borchers 2000, Haag 1992).

Positive Energy

In quantum physics the time development of the system is often represented by a strongly continuous group $\{U(t) = e^{itH} \mid t \in \mathbb{R}\}$ of unitary operators, and the generator H is interpreted as the total energy of the system. There is a link between modular structure and positive energy, which has found many applications in quantum field theory. This result was crucial in the development of Theorem 11 and was motivated by the 1975 discovery mentioned above, now commonly called the Bisognano-Wichmann theorem.

Theorem 14 Let \mathcal{M} be a von Neumann algebra with a cyclic and separating vector Ω , and let $\{U(t)\}$ be a continuous unitary group satisfying $U(t)\mathcal{M}U(-t) \subset \mathcal{M}$, for all $t \geq 0$. Then any two of the following conditions imply the third:

- (i) $U(t) = e^{itH}$, with $H \geq 0$;
- (ii) $U(t)\Omega = \Omega$, for all $t \in \mathbb{R}$; and
- (iii) $\Delta^{it}U(s)\Delta^{-it} = U(e^{-2\pi t}s)$ and $JU(s)J = U(-s)$, for all $s, t \in \mathbb{R}$.

Modular Nuclearity and Phase Space Properties

Modular theory can be used to express physically meaningful properties of quantum “phase spaces” by a condition of compactness or nuclearity of certain maps. In its initial form, the condition was formulated in terms of the Hamiltonian, the global energy operator of theories in Minkowski space. The above indications that the modular operators carry information about the energy of the system were reinforced when it was shown that a

formulation in terms of modular operators was essentially equivalent.

Let $\mathcal{O}_1 \subset \mathcal{O}_2$ be nonempty bounded open subregions of Minkowski space with corresponding algebras of observables $\mathcal{A}(\mathcal{O}_1) \subset \mathcal{A}(\mathcal{O}_2)$ in a vacuum representation with vacuum vector Ω , and let Δ be the modular operator associated with $(\mathcal{A}(\mathcal{O}_2), \Omega)$ (by the Reeh–Schlieder theorem, Ω is cyclic and separating for $\mathcal{A}(\mathcal{O}_2)$). For each $\lambda \in (0, 1/2)$ define the mapping $\Xi_\lambda: \mathcal{A}(\mathcal{O}_1) \rightarrow \mathcal{H}$ by $\Xi_\lambda(A) = \Delta^\lambda A \Omega$. The compactness of any one of these mappings implies the compactness of all of the others. Moreover, the l^p (nuclear) norms of these mappings are interrelated and provide a measure of the number of local degrees of freedom of the system. Suitable conditions on the maps in terms of these norms entail the strong statistical independence condition called the split property. Conversely, the split property implies the compactness of all of these maps. Moreover, the existence of equilibrium temperature states on the global algebra of observables can be derived from suitable conditions on these norms in the vacuum sector.

The conceptual advantage of the modular compactness and nuclearity conditions compared to their original Hamiltonian form lies in the fact that they are meaningful also for quantum systems in curved spacetimes, where global energy operators (i.e., generators corresponding to global timelike Killing vector fields) need not exist.

Modular Position and Quantum Field Theory

The characterization of the relative “geometric” position of algebras based on the notions of modular inclusion and modular intersection was directly motivated by the Bisognano–Wichmann theorem. Observable algebras associated with suitably chosen wedge regions in Minkowski space provided examples whose essential structure could be abstracted for more general application, resulting in the notions presented in the preceding sections.

Theorem 12(ii) has been used to construct, from two algebras and the indicated half-sided modular inclusions, a conformal quantum field theory on the circle (compactified light ray) with positive energy. Since the chiral part of a conformal quantum field model in two spacetime dimensions naturally yields such half-sided modular inclusions, studying the inclusions in Theorem 12(ii) is equivalent to studying such field theories. Theorems 12(i) and 13 and their generalizations to inclusions involving up to six algebras have been employed to construct Poincaré-covariant nets of observable algebras (the algebraic form of quantum field theories) satisfying the spectrum condition on $(d+1)$ -dimensional

Minkowski space for $d=1, 2, 3$. Conversely, such quantum field theories naturally yield such systems of algebras.

This intimate relation would seem to open up the possibility of constructing interacting quantum field theories from a limited number of modular inclusions/intersections.

Geometric Modular Action

The fact that the modular objects in quantum field theory associated with wedge-shaped regions and the vacuum state in Minkowski space have geometric significance (“geometric modular action”) was originally discovered in the framework of the Wightman axioms. As an algebraic quantum field theory (AQFT) does not rely on the concept of Wightman fields, it was natural to ask (i) when does geometric modular action hold in AQFT and (ii) which physically relevant consequences follow from this feature?

There are two approaches to the study of geometric modular action. In the first, attention is focused on modular covariance, expressed in terms of the modular groups associated with wedge algebras and the vacuum state in Minkowski space. Modular covariance has been proven to obtain in conformally invariant AQFT, in any massive theory satisfying asymptotic completeness, and also in the presence of other, physically natural assumptions. To mention only three of its consequences, both the spin–statistics theorem and the PCT theorem, as well as the existence of a continuous unitary representation of the Poincaré group acting covariantly upon the observable algebras and satisfying the spectrum condition follow from modular covariance.

In a second approach to geometric modular action, the modular involutions are the primary focus. Here, no *a priori* connection between the modular objects and isometries of the spacetime is assumed. The central assumption, given the state vector Ω and the von Neumann algebras of localized observables $\{\mathcal{A}(\mathcal{O})\}$ on the spacetime, is that there exists a family \mathcal{W} of subsets of the spacetime such that $J_{W_1} \mathcal{R}(W_2) J_{W_1} \in \{\mathcal{R}(W) \mid W \in \mathcal{W}\}$, for every $W_1, W_2 \in \mathcal{W}$. This condition makes no explicit appeal to isometries or other special attributes and is thus applicable, in principle, to quantum field theories on general curved spacetimes.

It has been shown for certain spacetimes, including Minkowski space, that under certain additional technical assumptions, the modular involutions encode enough information to determine the dynamics of the theory, the isometry group of the spacetime, and a continuous unitary representation of the isometry group which acts covariantly upon the observables and leaves the state invariant. In certain

cases including Minkowski space, it is even possible to derive the spacetime itself from the group \mathcal{J} generated by the modular involutions $\{J_W \mid W \in \mathcal{W}\}$.

The modular unitaries Δ_W^{it} enter in this approach through a condition which is designed to assure the stability of the theory, namely that $\Delta_W^{it} \in \mathcal{J}$, for all $t \in \mathbb{R}$ and $W \in \mathcal{W}$. In Minkowski space, this additional condition entails that the derived representation of the Poincaré group satisfies the spectrum condition.

Further Applications

As previously observed, through the close connection to the KMS condition, modular theory enters naturally into the equilibrium thermodynamics of many-body systems. But in recent work on the theory of nonequilibrium thermodynamics it also plays a role in making mathematical sense of the notion of quantum systems in local thermodynamic equilibrium. Modular theory has also proved to be of utility in recent developments in the theory of superselection rules and their attendant sectors, charges and charge-carrying fields.

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Quantum Central-Limit Theorems; Symmetries in Quantum Field Theory of

Lower Spacetime Dimensions; Thermal Quantum Field Theory; Positive Maps on C^* -Algebras; Two-Dimensional Models; von Neumann Algebras: Introduction, Modular Theory, and Classification Theory.

Further Reading

- Baumgärtel H and Wollenberg M (1992) *Causal Nets of Operator Algebras*. Berlin: Akademie-Verlag.
- Borchers HJ (2000) On revolutionizing quantum field theory with Tomita's modular theory. *Journal of Mathematical Physics* 41: 3604–3673.
- Bratteli O and Robinson DW (1981) *Operator Algebras and Quantum Statistical Mechanics II*. Berlin: Springer.
- Connes A (1974) Caractérisation des algèbres de von Neumann comme espaces vectoriels ordonnés. *Annales de l'Institut Fourier* 24: 121–155.
- Haag R (1992) *Local Quantum Physics*. Berlin: Springer.
- Kadison RV and Ringrose JR (1986) *Fundamentals of the Theory of Operator Algebras*, vol. II. Orlando: Academic Press.
- Pedersen GK (1979) *C^* -Algebras and Their Automorphism Groups*. New York: Academic Press.
- Stratila S (1981) *Modular Theory in Operator Algebras*. Tunbridge Wells: Abacus Press.
- Takesaki M (1970) *Tomita's Theory of Modular Hilbert Algebras and Its Applications*, Lecture Notes in Mathematics, vol. 128. Berlin: Springer.
- Takesaki M (2003) *Theory of Operator Algebras II*. Berlin: Springer.

Topological Defects and Their Homotopy Classification

T W B Kibble, Imperial College, London, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Symmetry-breaking phase transitions occur in a wide variety of systems – from condensed matter to the early universe. One of the common features of such transitions is the appearance, in the broken-symmetry phase, of topological defects, trapped regions in which the symmetry is restored, or at least changed. Examples are vortices in superfluids, domain walls in ferromagnets, and disclination lines in liquid crystals. Often these defects are stable for topological reasons, and play an important role in the dynamics of the system. An astonishingly rich variety of defects can be found in various systems. They can usefully be classified using the tools of homotopy theory.

Spontaneous Symmetry Breaking

Let us consider a quantum-mechanical system with a symmetry group G . This means that each $g \in G$ is

represented on the Hilbert space of quantum states by a unitary operator $\hat{U}(g)$, which commutes with the Hamiltonian. Spontaneous symmetry breaking occurs if this symmetry is not shared by the ground state or vacuum state $|0\rangle$ of the system. In other words, for some $g \in G$, $\hat{U}(g)|0\rangle \neq |0\rangle$. Then the ground state is necessarily degenerate: $\hat{U}(g)|0\rangle$ must have the same energy as $|0\rangle$.

Spontaneous symmetry breaking is usually describable in terms of an order-parameter field, which vanishes above the transition and is nonzero below it. We can find a scalar field $\hat{\phi}(\mathbf{r})$, or multiplet of fields $\hat{\phi} = (\hat{\phi}_i, i = 1, \dots, n)$ transforming according to some representation D of G (assumed not to contain the trivial representation), whose expectation value in the ground state is nonzero:

$$\langle 0 | \hat{\phi}(\mathbf{r}) | 0 \rangle = \phi_0 \neq 0 \quad [1]$$

This is the order parameter. Since

$$\langle 0 | \hat{U}^\dagger(g) \hat{\phi}(\mathbf{r}) \hat{U}(g) | 0 \rangle = D(g) \phi_0 \quad [2]$$

it follows that the only elements of G that can be symmetries of the ground state are those in the

stability subgroup H of ϕ_0 (the group of unbroken symmetries in this ground state):

$$H = \{g \in G: D(g)\phi_0 = \phi_0\} \quad [3]$$

In terms of this subgroup, we can find a useful characterization of the manifold \mathcal{M} of degenerate ground states. As noted above, for each $g \in G$, $\hat{U}(g)|0\rangle$ is also a ground state. However, these are not all distinct, because clearly $\hat{U}(gb)|0\rangle = \hat{U}(g)|0\rangle$ for all $b \in H$. Hence, the distinct ground states are in one-to-one correspondence with the left cosets gH of H in G , and \mathcal{M} may be identified with the quotient space G/H , the space of left cosets.

For example, suppose G is the rotation group $SO(3)$, and $\hat{\phi}$ belongs to the three-dimensional vector representation. If $\phi \neq 0$ in the ground state, we may choose $\phi_0 = (0, 0, v)$. Then, clearly, $H = SO(2)$, the group of rotations about the z -axis, and $\mathcal{M} = SO(3)/SO(2) = S^2$, the 2-sphere. It is useful to think of \mathcal{M} as the subset of the order-parameter space comprising the possible expectation values $\phi = \langle \hat{\phi} \rangle$ for the various degenerate ground states. For example, in this case, $\mathcal{M} = \{\phi: \phi^2 = v^2\}$.

Defect Formation

It is often possible to characterize the dynamics at finite temperature in terms of a function of the order parameter, the effective potential $V(\phi)$, which is necessarily invariant under G , and whose minima define the equilibrium states. At low temperatures, it has a form like $V = \lambda(\phi^2 - v^2)^2$, whose minima occur at nonzero values of ϕ . But above the critical temperature T_c , the only minimum is at $\phi = 0$, so the equilibrium state is symmetric under G . In the high-temperature phase, there may be large fluctuations in $\hat{\phi}$, but its mean value will be zero.

Now, when the system is cooled through the phase transition, $\hat{\phi}$ will acquire a nonzero expectation value, gradually approaching one of the degenerate ground states characterized by a point of \mathcal{M} . But the choice of which one is unpredictable; the symmetry breaking is spontaneous. Moreover, in a large system, there is no reason why the same choice should be made everywhere. For example, a ferromagnet cooling through its Curie point may acquire a spontaneous magnetization in different directions in different parts of the sample.

Of course, there is an energetic penalty to having a spatially varying order parameter, so it will tend to become more uniform as the temperature is lowered. But the question arises whether there may be any topological obstruction to this process. It can happen that if we choose points on \mathcal{M} in a

continuous manner everywhere around the periphery of some region, it is topologically impossible to complete the process throughout its interior. Continuity may require that there are points where ϕ leaves the surface \mathcal{M} . For example, if our ferromagnet has two opposite possible directions of easy magnetization, described by ϕ_0 and $-\phi_0$, then \mathcal{M} consists essentially of these two points. Regions where $\phi \approx \phi_0$ and where $\phi \approx -\phi_0$ must be separated by domain walls across which ϕ varies smoothly from one to the other.

Homotopy Groups

To classify the various possible types of defect, we need to consider the homotopy groups of the manifold \mathcal{M} of degenerate ground states. In this section, we briefly review the necessary definitions.

A path in \mathcal{M} is a map $\phi: I \rightarrow \mathcal{M}$ from the unit interval $I = [0, 1] \subset \mathbb{R}$. We choose a base point $m_0 \in \mathcal{M}$ (which may be identified with ϕ_0), and consider loops in \mathcal{M} , paths such that $\phi(0) = \phi(1) = m_0$. We say that two loops are homotopic, and write $\phi \sim \psi$, if one can be continuously deformed into the other within \mathcal{M} , that is, if there exists a map $\chi: I^2 \rightarrow \mathcal{M}$ such that

$$\chi(0, t) = \phi(t) \quad \text{and} \quad \chi(1, t) = \psi(t) \quad [4]$$

for all t , and

$$\chi(s, 0) = \chi(s, 1) = m_0 \quad [5]$$

for all s . This is an equivalence relation. The set $\pi_1(\mathcal{M})$ is the set of equivalence classes $[\phi]$ of loops under this relation.

On the set of loops, we may define a product $\phi\psi$, comprising the loop ϕ followed by ψ (see Figure 1). Explicitly,

$$(\phi\psi)(t) = \begin{cases} \phi(2t), & 0 \leq t \leq \frac{1}{2} \\ \psi(2t - 1), & \frac{1}{2} < t \leq 1 \end{cases} \quad [6]$$

It is easy to show that if $\phi \sim \phi'$ and $\psi \sim \psi'$, then $\phi\psi \sim \phi'\psi'$. Hence, this defines a product on $\pi_1(\mathcal{M})$, by $[\phi][\psi] = [\phi\psi]$. So equipped, $\pi_1(\mathcal{M})$ becomes the

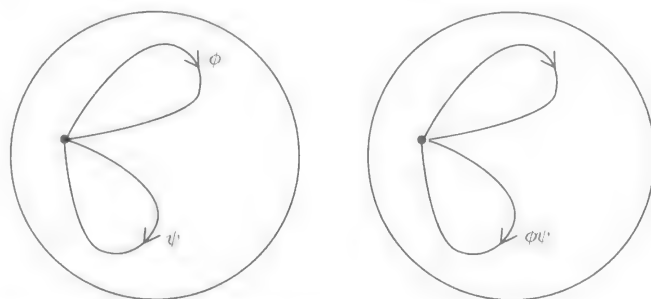


Figure 1 The product of loops.

fundamental group or first homotopy group of \mathcal{M} . Note that the identity is the equivalence class $[\phi_0]$ of the trivial loop with $\phi_0(t) \equiv m_0$, while the inverse is $[\phi]^{-1} = [\tilde{\phi}]$, where the map $\tilde{\phi}$ is the reverse of ϕ : $\tilde{\phi}(t) = \phi(1 - t)$.

Strictly speaking, we should write $\pi_1(\mathcal{M}, m_0)$ in place of $\pi_1(\mathcal{M})$. However, for any path-connected space, the groups $\pi_1(\mathcal{M}, m_0)$ and $\pi_1(\mathcal{M}, m'_0)$ are always isomorphic, and, more importantly, the same is true for any coset space $\mathcal{M} = G/H$, where G is a Lie group and H a closed subgroup. For a general manifold \mathcal{M} , $\pi_1(\mathcal{M})$ is not necessarily abelian, but it is so if \mathcal{M} is a Lie group, or more generally a Riemannian symmetric space. The space \mathcal{M} is said to be simply connected if $\pi_1(\mathcal{M}) = 0$, the group comprising only the identity element, $0 = \{[\phi_0]\}$. (Although $\pi_1(\mathcal{M})$ is not always abelian, it is conventional for homotopy groups to use an additive notation and represent the trivial group by 0 rather than 1.)

The n th homotopy group $\pi_n(\mathcal{M})$ may be defined similarly, as a set of equivalence classes of maps $\phi: I^n \rightarrow \mathcal{M}$ such that ϕ maps the entire boundary ∂I^n to the base point m_0 . Two such maps are homotopic ($\phi \sim \psi$) if there exists a map $\chi: I^{n+1} \rightarrow \mathcal{M}$ such that

$$\chi(0, t) = \phi(t) \quad \text{and} \quad \chi(1, t) = \psi(t) \quad [7]$$

for all $t = (t_1, \dots, t_n)$, and, for each $s \in I$, $\chi(s, t) = m_0$ for all $t \in \partial I^n$. The product $\phi\psi$ is defined by

$$(\phi\psi)(t_1, \dots, t_n) = \begin{cases} \phi(2t_1, t_2, \dots, t_n), & 0 \leq t_1 \leq \frac{1}{2} \\ \psi(2t_1 - 1, t_2, \dots, t_n), & \frac{1}{2} < t_1 \leq 1 \end{cases} \quad [8]$$

The choice of t_1 rather than any other t_j is arbitrary; all choices yield homotopic product maps. The product again defines a product on $\pi_n(\mathcal{M})$, which thereby becomes a group, the n th homotopy group. One new feature is that, for all $n > 1$, $\pi_n(\mathcal{M})$ is always abelian.

Note that since the entire boundary of I_n is mapped to a single point, it is possible to collapse it, and talk instead about maps from the n -sphere S^n to \mathcal{M} , taking one designated point to m_0 . The fact that $\pi_n(\mathcal{M})$ is nontrivial indicates the existence in \mathcal{M} of closed n -surfaces that cannot be smoothly shrunk to a point. In particular, it is worth noting that, for any n , $\pi_n(S^n) = \mathbb{Z}$, the additive group of integers, while $\pi_m(S^n) = 0$ for all $m < n$.

A special case is $n = 0$. Here, S^0 comprises two points only, and since one of them is always mapped to m_0 , we really have to consider maps from a single point to \mathcal{M} , that is, points in \mathcal{M} . Two points are homotopic if they can be joined by a path in \mathcal{M} . Thus, $\pi_0(\mathcal{M})$ may be identified with the set of path-connected components of \mathcal{M} . Note, however, that in

general no product can be defined on $\pi_0(\mathcal{M})$, so $\pi_0(\mathcal{M})$ should be called the zeroth homotopy set (not group). There is an important exception, however: if G is a Lie group, and G_0 its connected subgroup (the subset of elements joined by paths to the identity e), then $\pi_0(\mathcal{M})$ may be identified with the quotient group G/G_0 . Note, however, that this group $\pi_0(\mathcal{M}) = G/G_0$ is not necessarily abelian.

Classification of Defects

We now turn to the classification of defects by means of homotopy groups. It will be useful to start with simple specific examples in three-dimensional space, \mathbb{R}^3 .

First, suppose again that ϕ belongs to the vector representation of $G = \text{SO}(3)$. Then $\mathcal{M} = \text{SO}(3)/\text{SO}(2) = S^2$ may be identified with the sphere $\mathcal{M} = \{\phi: \phi^2 = v^2\}$ in ϕ space. Consider a closed surface \mathcal{S} , an embedding of a 2-sphere S^2 in \mathbb{R}^3 . Assume that everywhere on \mathcal{S} the field $\phi(\mathbf{r})$ has one of the ground-state values. In other words, we have a map $\phi: \mathcal{S} \rightarrow \mathcal{M}$, from one 2-sphere to another. The map ϕ can be extended to a map from the interior of \mathcal{S} to \mathcal{M} only if it belongs to the trivial homotopy class $[\phi_0] \in \pi_2(\mathcal{M})$, where $\phi_0: I^2 \rightarrow \mathcal{M}: (t_1, t_2) \mapsto m_0 = eH$. In all other cases, there must be at least one point where $\phi(\mathbf{r}) = 0$; this is a point defect. The second homotopy group in this case is $\pi_2(S^2) = \mathbb{Z}$, so the possible point defects, or monopoles, are labeled by an integer $n \in \mathbb{Z}$, the winding number. (An example of a map with winding number n is (in spherical polars) $(r, \theta, \varphi) \mapsto (v, \theta, n\varphi)$.)

More generally, point defects in \mathbb{R}^d are classified by $\pi_{d-1}(\mathcal{M})$. A map ϕ from a closed $(d-1)$ -dimensional surface $\mathcal{S} \subset \mathbb{R}^d$ to \mathcal{M} can be extended to the interior of \mathcal{S} if and only if it belongs to the trivial homotopy class $[\phi_0] \in \pi_{d-1}(\mathcal{M})$. If this is not the case, there must be at least one point around which $\phi(\mathbf{r})$ leaves the surface \mathcal{M} , although in general it is not required to vanish anywhere.

Second, take the case where ϕ is a single complex field, and G is the phase symmetry group $\text{U}(1)$. In this case, H is the subgroup $1 = \{1\} \subset G$. Thus, $\mathcal{M} = \text{U}(1)/1 = S^1$; this manifold may be identified with the circle $\{\phi: |\phi| = v\}$ in the order-parameter space. Now consider a closed loop \mathcal{C} in space, an embedding of S^1 in \mathbb{R}^3 (see Figure 2). Suppose that on \mathcal{C} , $\phi(\mathbf{r})$ takes one of the ground-state values, say $\phi(\mathbf{r}) = v \exp[i\alpha(\mathbf{r})]$. If \mathcal{S} is some surface with boundary \mathcal{C} , then the map $\phi: \mathcal{C} \rightarrow \mathcal{M}$ can be extended to a map $\phi: \mathcal{S} \rightarrow \mathcal{M}$ if and only if it belongs to the trivial homotopy class $[\phi_0] \in \pi_1(\mathcal{M})$. If it does not, then there must be at least one point on \mathcal{S} within \mathcal{C} where $\phi = 0$. Moreover, this

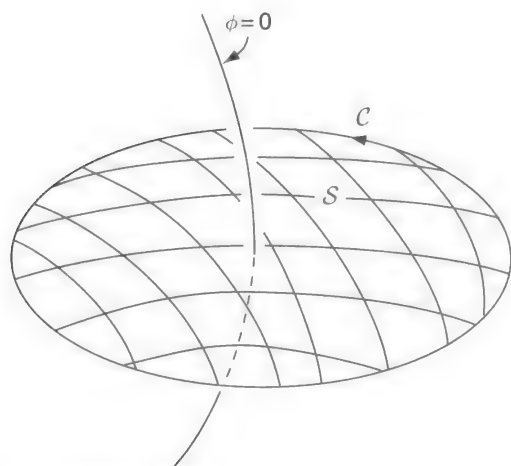


Figure 2 A linear defect.

must be true of every surface S spanning C , so there must be a curve passing through C along which $\phi=0$. This is a linear defect, a string or vortex line. In this case, the first homotopy group is $\pi_1(S^1)=\mathbb{Z}$, so we see that the possible linear defects are classified by an integer, the winding number n . An example of a map with winding number n is $\varphi \mapsto ve^{ni\varphi}$.

Again, this result can easily be generalized. Linear defects in \mathbb{R}^d are classified by $\pi_{d-2}(\mathcal{M})$. If, on a $(d-2)$ -dimensional surface C , $\phi(r)$ takes values in \mathcal{M} , and if it does not belong to the trivial homotopy class, there must be a linear defect threading through C , around which ϕ leaves the surface \mathcal{M} – although again it need not necessarily vanish.

More generally yet, in the d -dimensional space \mathbb{R}^d , defects of dimension p are classified by the homotopy group $\pi_{d-p-1}(\mathcal{M})$. For example, in three dimensions, planar defects – domain walls – are classified by $\pi_0(\mathcal{M})$.

The Exact Sequence

There are mathematical theorems that greatly facilitate the computation of the homotopy group of homogeneous spaces, of the form $\mathcal{M}=G/H$.

We begin with the maps relating these spaces to each other. There is a canonical injective homomorphism $i:H \rightarrow G:h \mapsto h$, and a canonical projection associating each element of G with its coset: $p:G \rightarrow \mathcal{M}:g \mapsto gH$. Moreover, it is clear that the image of i , namely the subgroup H , is also the kernel of p , the inverse image $p^{-1}m_0$ of the distinguished element $m_0=eH$ of \mathcal{M} . These statements can be summarized by saying that

$$1 \rightarrow H \xrightarrow{i} G \xrightarrow{p} \mathcal{M} \rightarrow 1$$

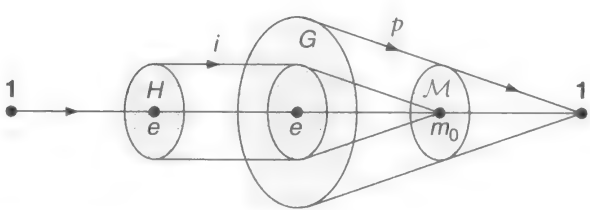


Figure 3 An exact sequence.

is an exact sequence: the image of each map is the kernel of the following one (see Figure 3).

Next, we note that since any closed loops (or n -surfaces) in H belonging to the same homotopy class are also homotopic as loops (or n -surfaces) in G , there is an induced homomorphism $i_*:\pi_n(H) \rightarrow \pi_n(G)$. Similarly, homotopic loops or n -surfaces in G project to homotopic loops or n -surfaces in \mathcal{M} , so there is an induced homomorphism $p_*:\pi_n(G) \rightarrow \pi_n(\mathcal{M})$. Moreover, it is easy to see that although i_* is not necessarily injective and p_* not necessarily projective, it is true that the image of i_* is the kernel of p_* . For example, any loop in G will be mapped to a homotopically trivial loop in \mathcal{M} if and only if it is homotopic to the image of a loop in H .

In addition, there is a boundary map that relates homotopy groups of different dimension: $\partial:\pi_{n+1}(\mathcal{M}) \rightarrow \pi_n(H)$. To see this, it is useful to think of G as a fiber bundle with base space \mathcal{M} and fiber H . Now consider a map $\phi:(I^{n+1}, \partial I^{n+1}) \rightarrow (\mathcal{M}, m_0)$. Since p is a projection, ϕ can always be lifted to a map $\hat{\phi}:(I^{n+1}, \partial I^{n+1}) \rightarrow (G, H)$, that is, we can find a (nonunique) map $\hat{\phi}$ such that $\phi=p \circ \hat{\phi}$ (see Figure 4). However, $\hat{\phi}$ does not necessarily map the boundary to a single point; what is true is that $\hat{\phi}$ must map the boundary to a subset of H , and since topologically $\partial I^{n+1} \simeq S^n$, this defines a map $\tilde{\phi}:S^n \rightarrow H$. If we allow ϕ to vary over some homotopy class of maps, and $\hat{\phi}$ to vary continuously, then $\tilde{\phi}$ will

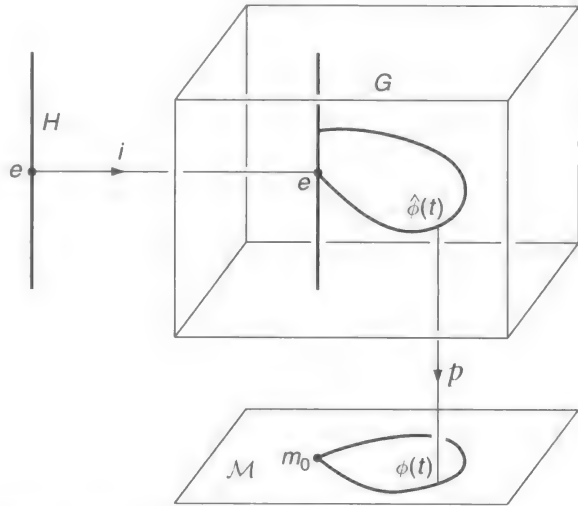


Figure 4 Lift of a loop.

also remain in one homotopy class. Thus, we have defined a map $\partial: \pi_{n+1}(\mathcal{M}) \rightarrow \pi_n(H): [\phi] \mapsto [\tilde{\phi}]$.

It is also easy to see that the image of $\partial: \pi_{n+1}(\mathcal{M}) \rightarrow \pi_n(H)$ is the kernel of $i_*: \pi_n(H) \rightarrow \pi_n(G)$, because the n -surface in H defined by $\tilde{\phi}$ is necessarily homotopically trivial in G . Similarly, one can see that the image of $p_*: \pi_{n+1}(G) \rightarrow \pi_{n+1}(\mathcal{M})$ is the kernel of $\partial: \pi_{n+1}(\mathcal{M}) \rightarrow \pi_n(H)$.

Putting all these results together, we see that there is a (semi-infinite) exact sequence connecting all the homotopy groups:

$$\begin{aligned} \cdots \xrightarrow{p_*} \pi_{n+1}(\mathcal{M}) \xrightarrow{\partial} \pi_n(H) \xrightarrow{i_*} \pi_n(G) \xrightarrow{p_*} \pi_n(\mathcal{M}) \\ \xrightarrow{\partial} \pi_{n-1}(H) \xrightarrow{i_*} \cdots \pi_1(G) \xrightarrow{p_*} \pi_1(\mathcal{M}) \xrightarrow{\partial} \pi_0(H) \quad [9] \\ \xrightarrow{i_*} \pi_0(G) \xrightarrow{p_*} \pi_0(\mathcal{M}) \end{aligned}$$

This sequence makes it easy to compute most of the low-dimensional homotopy groups of \mathcal{M} . Let us begin with $\pi_0(\mathcal{M})$, which merely labels its disconnected components. As noted earlier, for the Lie group G , $\pi_0(G)$ is the quotient group $\pi_0(G) = G/G_0$, where G_0 is the connected subgroup of G . Now the image of $\pi_0(H)$ under i_* is clearly the set of connected components of G that contain elements of H , so if G has m connected components, and n of them contain elements of H , then $\pi_0(\mathcal{M})$ has m/n elements (see Figure 5).

Next, we note that, for all the higher homotopy groups, disconnected pieces are irrelevant. Since a loop, for example, starting at m_0 must remain within its connected component $\mathcal{M}_0 \subset \mathcal{M}$, it follows that $\pi_1(\mathcal{M}) = \pi_1(\mathcal{M}_0)$, and similarly $\pi_n(\mathcal{M}) = \pi_n(\mathcal{M}_0)$ for all $n > 1$. So one can ignore any disconnected parts of the symmetry group G , and assume from now on that $\pi_0(G) = 0$. Moreover, it is always possible to replace G by its simply connected covering group, replacing $\text{SO}(3)$, for

example, by $\text{SU}(2)$. Thus, we may also assume that $\pi_1(G) = 0$. Then the section of the exact sequence in the second line of [9] becomes

$$0 \xrightarrow{p_*} \pi_1(\mathcal{M}) \xrightarrow{\partial} \pi_0(H) \xrightarrow{i_*} 0$$

which implies that the two groups in the center are isomorphic:

$$\pi_1(\mathcal{M}) = \pi_0(H) \quad [10]$$

For example, if the symmetry group $G = \text{SO}(3)$ is completely broken, so that $H = 1$, then replacing G by $\tilde{G} = \text{SU}(2)$ requires replacing H by $\tilde{H} = \{+1, -1\} \simeq \mathbb{Z}_2$, hence also $\pi_1(\mathcal{M}) = \pi_0(\tilde{H}) = \mathbb{Z}_2$; there is only one nontrivial class of linear defects in this model.

To find $\pi_2(\mathcal{M})$, we need a standard theorem about Lie groups, namely that the second homotopy group of any Lie group is trivial: for any G , $\pi_2(G) = 0$. (No details of the proof are given here. It derives from the fact that a generic element $g \in G$ belongs to a unique one-parameter subgroup $\{\exp(tX), t \in \mathbb{R}\} \subset G$, where X is an element of the Lie algebra of G . Thus, all the points on a surface in G may be joined by these paths to the identity, and the surface may then be shrunk along the resulting cone. There are exceptional elements for which this is not true, but it can be shown that in a d -dimensional group they lie on $(d-3)$ -dimensional surfaces, so any 2-surface can be smoothly deformed to avoid them.)

It follows from this theorem that another section of the exact sequence is

$$0 \xrightarrow{p_*} \pi_2(\mathcal{M}) \xrightarrow{\partial} \pi_1(H) \xrightarrow{i_*} 0$$

which again implies an isomorphism:

$$\pi_2(\mathcal{M}) = \pi_1(H) \quad [11]$$

For example, if $G = \text{SO}(3)$ and $H = \text{SO}(2)$, or equivalently $\tilde{G} = \text{SU}(2)$ and $\tilde{H} = \text{U}(1)$ (a double cover of the $\text{SO}(2)$), then $\pi_2(\mathcal{M}) = \pi_1(\tilde{H}) = \mathbb{Z}$, so point defects in this theory are labeled by an integer winding number.

Examples

The simplest continuous symmetry is the $\text{U}(1)$ phase symmetry $\hat{\phi} \mapsto \hat{\phi}e^{i\alpha}$ of a complex field. In a weakly interacting Bose gas, below the Bose-Einstein condensation temperature, or in superfluid helium-4, a macroscopic fraction of the atoms occupies a single quantum state, and $\hat{\phi}$ acquires a nonzero expectation value, $\langle \hat{\phi} \rangle = \phi$, whose phase is arbitrary, so the symmetry is completely broken to $H = 1$. Thus, $\mathcal{M} = S^1$; we have a circle of equivalent degenerate ground states. (This corresponds to

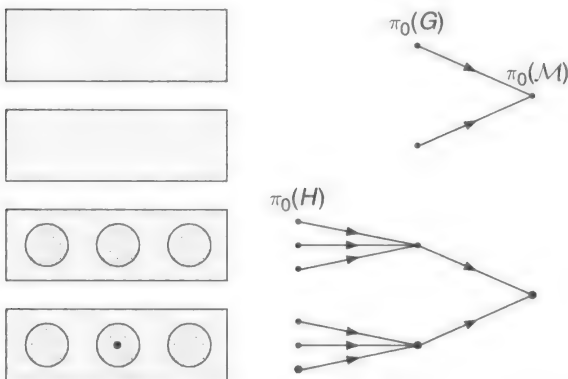


Figure 5 The disconnected components of G are shaded, those of H are cross-hatched. Here $\pi_0(\mathcal{M})$ has two elements.

spontaneous breaking of the particle-number symmetry. It is possible to describe the system in a $U(1)$ -invariant way, by projecting out a state of definite particle number, a uniform superposition of all the states in \mathcal{M} , but it is generally less convenient to do so.) In this case, the only nontrivial homotopy group is $\pi_1(\mathcal{M}) = \mathbb{Z}$, so the only defects are linear defects classified by a winding number $n \in \mathbb{Z}$. The defects with $n = \pm 1$ are stable vortices. Those with $|n| > 1$ are in general unstable and tend to break up into $|n|$ single-quantum vortices.

Low-temperature superconductors also have a $U(1)$ symmetry, although there are important differences. This is not a global symmetry but a local, gauge symmetry, with coupling to the electromagnetic field. Moreover, it is not single atoms that condense but Cooper pairs, pairs of electrons of equal and opposite momentum and spin. These systems too exhibit linear defects, magnetic flux tubes carrying a magnetic flux $4\pi n\hbar/e$.

A less trivial example is a nematic liquid crystal. These materials are composed of rod-shaped molecules that tend, at low temperatures, to line up parallel to one another. The nematic state is characterized by a preferred orientation, described by a unit vector \mathbf{n} , the director. (Note that \mathbf{n} and $-\mathbf{n}$ are physically equivalent.) There is long-range orientational order, with molecules preferentially lining up parallel to \mathbf{n} , but unlike a solid crystal there is no long-range translational order – the molecules move freely past each other as in a normal liquid.

A convenient order parameter here is the mean mass quadrupole tensor Φ of a molecule. In the nematic state, Φ is proportional to $(3\mathbf{n}\mathbf{n} - 1)$; for example, if $\mathbf{n} = (0, 0, 1)$, then Φ is diagonal with diagonal elements proportional to $(-1, -1, 2)$. In this case, the symmetry group is $SO(3)$ (or, more precisely, $O(3)$; but the inversion symmetry is not broken, so we can restrict our attention to the connected part of the group). The subgroup H that leaves this Φ invariant is a semidirect product, $H = SO(2) \ltimes \mathbb{Z}_2$ (isomorphic to $O(2)$), composed of rotations about the z -axis and rotations through π about axes in the x - y plane. (If we enlarge G to its simply connected covering group $\tilde{G} = SU(2)$, then H becomes $\tilde{H} = [U(1) \ltimes \mathbb{Z}_4]/\mathbb{Z}_2$, where $U(1)$ is generated as before by J_z . The essential difference is that the square of any of the elements in the disconnected piece of \tilde{H} is not now the identity but the element $e^{2\pi i J_z} = -1 \in U(1)$.) The manifold \mathcal{M} of degenerate ground states in this case is the projective space \mathbb{RP}^2 (obtained by identifying opposite points of S^2).

Since \tilde{H} has disconnected pieces, we have $\pi_1(\mathcal{M}) = \pi_0(\tilde{H}) = \mathbb{Z}_2$. Thus, there can be topologically

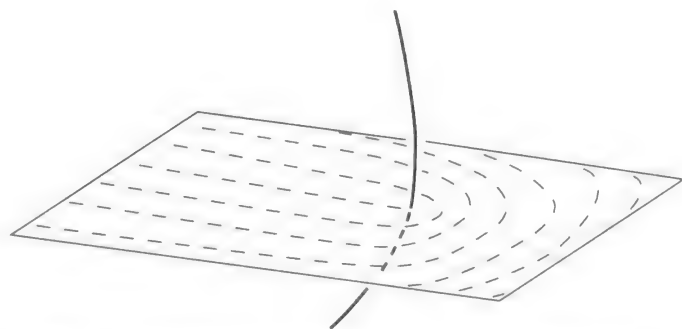


Figure 6 Orientation of molecules around a disclination line.

stable linear defects, here called disclination lines, around which the director \mathbf{n} rotates by π (see Figure 6). The fact that these defects are classified by \mathbb{Z}_2 rather than \mathbb{Z} means that a line around which \mathbf{n} rotates by 2π is topologically trivial; indeed, \mathbf{n} can be smoothly rotated near the line to run parallel to it, leaving a configuration with no defect.

There are also point defects; since $\pi_2(\mathcal{M}) = \pi_1(\tilde{H}) = \mathbb{Z}$, they are labeled by an integer winding number n . In a defect with $n = 1$, the vector \mathbf{n} points radially outwards all round the defect position.

Helium-3

Finally, let us turn to helium-3, one of the most fascinating and complex examples of spontaneous symmetry breaking, which becomes a superfluid at a temperature of a few millikelvin. Unlike helium-4, this is, of course, a Fermi liquid, so it is not the atoms that condense, but bound pairs of atoms, analogous to Cooper pairs. In this case, however, the most attractive channel is not the 1S , but the 3P , so the pairs have both orbital and spin angular momentum, $L = S = 1$. Therefore, the order parameter is not a single complex scalar field but a 3×3 complex matrix Φ_{jk} , where the two indices label the orbital and spin angular momentum states.

To a good approximation, the system is invariant under separate rotations of L and S (the effects of the small spin-orbit coupling will be discussed later), so the symmetry group is

$$G = U(1)_Y \times SO(3)_L \times SO(3)_S \quad [12]$$

where the subscripts denote the generators and $U(1)_Y$ represents multiplication by an overall phase factor, $e^{i\alpha Y} : \Phi_{jk} \mapsto \Phi_{jk} e^{i\alpha}$. This complicated symmetry allows much scope for a large variety of defects. There are, in fact, two distinct superfluid phases, A and B , with different symmetries (and indeed in the presence of a magnetic field there is a third, $A1$).

In the $^3\text{He-A}$ phase, the order parameter has the form $\Phi_{jk} \propto (m_j + in_j)d_k$, where \mathbf{m} , \mathbf{n} , \mathbf{d} are unit vectors, with $\mathbf{m} \perp \mathbf{n}$; if we set $\mathbf{l} = \mathbf{m} \wedge \mathbf{n}$, then

l defines the orbital angular momentum state by $l \cdot L = 1$, while d defines the spin quantization axis, such that $d \cdot S = 0$. The manifold \mathcal{M}_A for this phase is

$$\mathcal{M}_A = [\text{SO}(3) \times S^2]/\mathbb{Z}_2 \quad [13]$$

where the \mathbb{Z}_2 is present because (m, n, d) and $(-m, -n, -d)$ represent the same state. If, for example, we take l and d in the z -direction, the unbroken symmetry subgroup is

$$H_A = \text{SO}(2)_{L_z+Y} \times [\text{SO}(2)_{S_z} \ltimes \mathbb{Z}_2] \quad [14]$$

where the nontrivial element of \mathbb{Z}_2 may be taken to be $e^{i\pi(S_x+L_z)}$. The covering group of G is, of course,

$$\tilde{G} = \mathbb{R}_Y \times \text{SU}(2)_L \times \text{SU}(2)_S \quad [15]$$

Correspondingly,

$$\tilde{H}_A = \mathbb{R}_{L_z+Y} \times [\text{U}(1)_{S_z} \ltimes \mathbb{Z}_4] \quad [16]$$

It follows that the homotopy groups are

$$\pi_0(\mathcal{M}_A) = 0, \quad \pi_1(\mathcal{M}_A) = \mathbb{Z}_4, \quad \pi_2(\mathcal{M}_A) = \mathbb{Z} \quad [17]$$

There are linear defects labeled by a mod-4 quantum number and point defects labeled by an integer.

For the $^3\text{He-B}$ phase, by contrast, the order parameter is of the form

$$\Phi_{jk} \propto R_{jk} e^{i\theta} \quad [18]$$

where R is a rotation matrix, $R \in \text{SO}(3)$. Here then,

$$\mathcal{M}_B = \text{SO}(3) \times S^1 \quad [19]$$

with homotopy groups

$$\begin{aligned} \pi_0(\mathcal{M}_B) &= 0, \quad \pi_1(\mathcal{M}_B) = \mathbb{Z}_2 \times \mathbb{Z}, \\ \pi_2(\mathcal{M}_B) &= 0 \end{aligned} \quad [20]$$

In this phase, there are two distinct types of linear defect, the mass vortices with an integer label, and the spin vortices with a mod-2 label. (One can also have a “spin-mass vortex” carrying both quantum numbers.)

Composite Defects

There are several cases, including in particular helium-3, that exhibit symmetry breaking with multiple length or energy scales. For example, there may be two order parameters, say ϕ, ψ , with $|\phi| \gg |\psi|$. If $|\psi|$ is negligible, the symmetry G is broken by ϕ to H , and the manifold of degenerate ground states is $\mathcal{M} = G/H$. However, these states are not all exactly degenerate: ψ breaks the symmetry further to $K \subset H$, so the precisely degenerate ground states form a submanifold $\mathcal{M}' = G/K$.

The case of helium-3 is slightly different. Here it is the small spin-orbit coupling, arising from long-range dipole-dipole interactions, that introduces the second scale. Its effect is only significant over large distances.

In the $^3\text{He-A}$ phase, at short range the l and d vectors are uncorrelated but, over large distances, they tend to be aligned parallel or antiparallel. We can use the \mathbb{Z}_2 symmetry mentioned earlier to choose $l = d$. Hence, the manifold \mathcal{M}'_A of true ground states is only a submanifold of \mathcal{M}_A , namely $\mathcal{M}'_A = \text{SO}(3)$, whose homotopy groups are

$$\pi_0(\mathcal{M}'_A) = 0, \quad \pi_1(\mathcal{M}'_A) = \mathbb{Z}_2, \quad \pi_2(\mathcal{M}'_A) = 0 \quad [21]$$

Because of different behavior on different scales, “composite” defects can arise. For example, because $\pi_2(\mathcal{M}_A) = \mathbb{Z}$, there are short-range monopole configurations. For the $n=1$ monopole, we have a configuration with uniform l , and with d pointing outwards from the center. But, eventually the misalignment of d with l is energetically disfavored, and at large distances d tends to rotate to align with l except around one particular direction where it is oppositely aligned (see Figure 7). We have a composite defect: a small monopole coupled to a relatively fat string.

To see how the small- and large-scale structures fit together, one has to look also at the relative homotopy groups $\pi_n(\mathcal{M}, \mathcal{M}')$, whose elements are homotopy classes of maps from I^n to \mathcal{M} such that one face of the boundary is mapped into \mathcal{M}' , and the remainder to the chosen base point m_0 . For example, $\pi_1(\mathcal{M}, \mathcal{M}')$ classifies paths that terminate at m_0 while beginning at any point of \mathcal{M}' . There is, in fact, a long exact sequence, similar to [9], relating these homotopy groups, of which a typical segment is

$$\begin{aligned} \cdots \xrightarrow{\partial} \pi_n(\mathcal{M}') \xrightarrow{i_n} \pi_n(\mathcal{M}) \xrightarrow{p_n} \pi_n(\mathcal{M}, \mathcal{M}') \\ \xrightarrow{\partial} \pi_{n-1}(\mathcal{M}') \xrightarrow{i_{n-1}} \cdots \end{aligned} \quad [22]$$

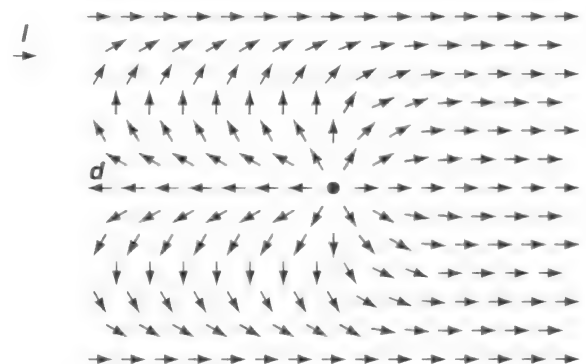


Figure 7 Cross-section of a short-range monopole attached to a fat string.

The relevant groups in the present case are

$$\pi_1(\mathcal{M}_A, \mathcal{M}'_A) = \mathbb{Z}_2, \quad \pi_2(\mathcal{M}_A, \mathcal{M}'_A) = \mathbb{Z} \quad [23]$$

Because $\pi_1(\mathcal{M}_A) = \mathbb{Z}_4$, there are three distinct classes of linear defects at small scales, but only those with quantum number $n = 2 \pmod{4}$ survive unchanged to large scales; they correspond to the nontrivial element of $\pi_1(\mathcal{M}'_A) = \mathbb{Z}_2$. On the other hand, the homotopy classes $n = \pm 1 \pmod{4}$ are mapped to nontrivial elements of $\pi_1(\mathcal{M}_A, \mathcal{M}'_A) = \mathbb{Z}_2$, which indicates that the corresponding linear defects are coupled at long range to fat domain walls, across which d rotates through π with a compensating rotation through π about l . Similarly, the nontrivial elements of $\pi_2(\mathcal{M}_A) = \mathbb{Z}$ are mapped to nontrivial elements of $\pi_2(\mathcal{M}_A, \mathcal{M}'_A)$, confirming that these short-range monopoles are coupled to fat strings, as in Figure 7.

For $^3\text{He-B}$, the effect of the spin-orbit coupling is to make the most energetically favorable configurations those in which the rotation matrix R in [18] represents a rotation about an arbitrary axis n through the Leggett angle $\theta_L = \arccos(-1/4) = 104^\circ$: $R = \exp(-i\theta_L n \cdot J)$.

Consequently,

$$\mathcal{M}'_B = S^2 \times S^1 \quad [24]$$

and so

$$\pi_0(\mathcal{M}'_B) = 0, \quad \pi_1(\mathcal{M}'_B) = \mathbb{Z}, \quad \pi_2(\mathcal{M}'_B) = \mathbb{Z} \quad [25]$$

The relative homotopy groups are

$$\pi_1(\mathcal{M}_B, \mathcal{M}'_B) = \mathbb{Z}_2, \quad \pi_2(\mathcal{M}_B, \mathcal{M}'_B) = 0 \quad [26]$$

Here the mass vortex persists at long range, but the configuration around the spin vortex deforms so that they become attached to fat domain walls. The “monopole” configurations corresponding to

nontrivial elements of $\pi_2(\mathcal{M}'_B)$ have no short-range singularity at all.

See also: Abelian Higgs Vortices; Leray–Schauder Theory and Mapping Degree; Liquid Crystals; Phase Transition Dynamics; Quantum Field Theory: A Brief Introduction; Quantum Fields with Topological Defects; Solitons and Other Extended Field Configurations; String Topology: Homotopy and Geometric Perspectives; Symmetries and Conservation Laws; Symmetry Breaking in Field Theory; Variational Techniques for Ginzburg–Landau Energies.

Further Reading

Helgason S (2001) *Differential Geometry, Lie Groups and Symmetric Spaces*. Providence, RI: American Mathematical Society.

Hu S-T (1959) *Homotopy Theory*. New York: Academic Press.

Kibble TWB (1976) Topology of cosmic domains and strings. *Journal of Physics A: Mathematical and General* 9: 1387–1398.

Kibble TWB (2000) Classification of topological defects and their relevance to cosmology and elsewhere. In: Bunkov YM and Godfrin H (eds.) *Topological Defects and the Non-Equilibrium-Dynamics of Symmetry Breaking Phase Transitions*, NATO Science Series C: Mathematical and Physical Sciences. vol. 249, pp. 7–31. Dordrecht: Kluwer Academic Publishers.

Shellard EPS and Vilenkin A (1994) *Cosmic Strings and Other Topological Defects*. Cambridge: Cambridge University Press.

Tilley DR and Tilley J (1990) *Superfluidity and Superconductivity*, 3rd edn. Bristol: IoP Publishing.

Toulouse G and Kléman M (1976) Principles of a classification of defects in ordered media. *Journal de Physique Lettres* 37: 149.

Vollhardt D and Wölfle P (1990) *The Superfluid Phases of Helium 3*. London: Taylor and Francis.

Volovik GE (1992) *Exotic Properties of Superfluid ^3He* . Singapore: World Scientific.

Volovik GE and Mineev VP (1977) Investigation of singularities in superfluid ^3He and in liquid crystals by the homotopic topology methods. *Zhurnal Eksperimentalnoi i Teoreticheskoi Fiziki* 72: 2256–2274 (*Soviet Physics–JETP* 24: 1186–1196).

Topological Gravity, Two-Dimensional

T Eguchi, University of Tokyo, Tokyo, Japan
© 2006 Elsevier Ltd. All rights reserved.

Introduction

It is well known that large- N Hermitian matrix models generate Feynman diagrams which represent the triangulation of Riemann surfaces. For instance, if we consider the integral of an $N \times N$ Hermitian matrix H

$$Z = \int dH \exp\left(-N\left[\frac{1}{2}\text{tr} H^2 + \frac{\lambda}{4}\text{tr} H^4\right]\right) \quad [1]$$

we find that the free energy $F = \log Z$ has the $1/N$ expansion

$$F = \sum_{g=0}^{\infty} N^{2-2g} F_g(\lambda) \quad [2]$$

Inspection of the Feynman diagrams shows that F_g reproduces the sum over the triangulations of genus g Riemann surfaces. The theory [1] is obviously well defined for $\lambda \geq 0$. In the large- N expansion, the theory continues to exist also at negative values of λ down to the critical point $\lambda_c = -1/12$.

The double scaling limit of large- N matrix models (Brézin and Kazakov 1990, Douglas and

Shenker 1990, Gross and Migdal 1990) is given by adjusting the coupling λ to λ_c and at the same time taking the limit $N \rightarrow \infty$. In this limit, contributions of all genera survive, and the theory describes the dynamics of fluctuating surfaces of arbitrary topologies. Results obtained in this way do not, in fact, depend on the detailed choice of the potential (ϕ^4 type in [1]) and have a high degree of universality. Thus, it provides an interesting model of two-dimensional (2D) quantum gravity.

Soon after the discovery of double scaling limit of matrix models, Witten observed that the correlation functions of the 2D gravity theory may be given a geometrical interpretation as topological invariants of the moduli space of Riemann surfaces \mathcal{M} , and that the 2D gravity theory may be reformulated as a topological field theory (Witten 1990). This reformulation of the results of the 2D gravity theory is called “2D topological gravity.”

In fact, 2D gravity theories come in a family parametrized by a pair of integers (p, q) . The double scaling limit of [1] gives the simplest example ($p = 2, q = 1$). Models with a chain of $p - 1$ Hermitian matrices give the (p, q) 2D gravity theories. The label q stands for the order of criticality of the model, and higher values of q are achieved by fine-tuning the parameters of the potential. At $q = 1$, 2D gravity theories possess a topological interpretation. The most basic case ($p = 2, q = 1$) is called pure topological gravity, and in theories at higher values of p , topological gravity is coupled to a matter system, that is, topological minimal models. Topological minimal models are obtained by twisting $\mathcal{N} = 2$ superconformal field theories.

Let us first consider the case of pure gravity ($p = 2, 1$). Let \mathcal{O}_n denote the observables in the theory and t_n the coupling constants to these operators. The correlation functions of topological gravity are given by

$$\langle \mathcal{O}_{n_1} \mathcal{O}_{n_2} \dots \mathcal{O}_{n_s} \rangle_g, \quad n_i = 1, 2, \dots \tag{3}$$

where $\langle \dots \rangle_g$ denotes the expectation value on a surface with g handles. The precise significance of eqn [3] as the intersection number on the moduli space is discussed below. The string partition function $\tau(t)$ is defined as the generating function of all possible correlation functions

$$\tau(t) = \exp \sum_{g=0}^{\infty} \left\langle \exp \sum t_n \mathcal{O}_n \right\rangle_g \tag{4}$$

The most striking aspect of topological gravity is the connection of the intersection theory on \mathcal{M} to the theory of completely integrable systems, that is, Kortevég–de Vries (KdV) and KP hierarchies. Witten conjectured that the generating function of

intersection numbers on moduli space $\tau(t)$ is the τ -function of KdV hierarchy. KdV hierarchy is obtained by generalizing the well-known KdV equation

$$\frac{\partial u}{\partial t} = \frac{3}{2} u \frac{\partial u}{\partial x} + \frac{1}{4} \frac{\partial^3 u}{\partial x^3} \tag{5}$$

Identification of the KdV equation with topological gravity is given by $u = 2 \langle \mathcal{O}_1 \mathcal{O}_1 \rangle, x = t_1, t = t_3$. Witten’s conjecture was verified by Kontsevich (1991) by an explicit construction of a new type of matrix model which generates the triangulation of the moduli space of Riemann surfaces.

In the general case of $(p, 1)$ topological gravity, the partition function of the theory obeys the equations of p th generalized KdV hierarchy (p reduction of KP hierarchy).

Intersection Theory

We now present some basic features of intersection theory on the moduli space of Riemann surfaces. It is known that 2D oriented surfaces Σ with g handles and s marked points x_i ($i = 1, \dots, s$) possess a finite number of inequivalent complex structures (complex structures are identified when they differ only by diffeomorphism). The space of inequivalent complex structures is called the moduli space $\mathcal{M}_{g,s}$ of the Riemann surface Σ . Its dimension is given by

$$\dim \mathcal{M}_{g,s} = 3g - 3 + s \tag{6}$$

For a mathematically rigorous treatment, we have to consider a compactification $\bar{\mathcal{M}}_{g,s}$ of moduli space $\mathcal{M}_{g,s}$ by adding suitable boundary components which arise due to various types of degenerations of Riemann surfaces. In the Deligne–Mumford or stable compactification, one considers the following three classes of singular Riemann surfaces Σ :

- 1. Two points, x_i and x_j , on Σ come close together. In this case, an extra 2-sphere is pinched off from the surface by forming a thin neck. The sphere contains points x_i and x_j and also the point x_l at the end of the neck (see Figure 1a). Since the original surface now has one point less and the 2-sphere with three points has no moduli, the degenerate surface has $3g - 4 + s$ parameters and forms a boundary divisor of the moduli space $\bar{\mathcal{M}}_{g,s}$.
- 2. If a cycle of nontrivial homology class shrinks to a point, we have a surface with one less genus and two extra marked points. Singular surface has $3(g - 1) - 3 + s + 2$ number of moduli and this is again a complex codimension-1 component (see Figure 1b).

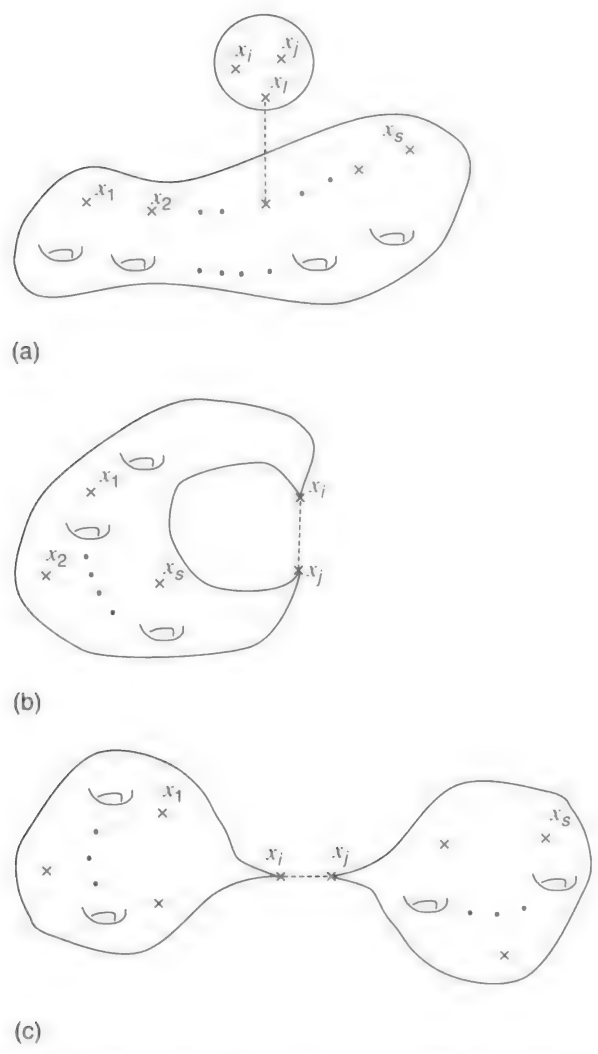


Figure 1 Degenerate Riemann surface obtained when (a) the points x_i and x_j coincide; (b) a nontrivial cycle collapses, two new points x_i and x_j are created; (c) a pinching cycle collapses, two new points x_i and x_j are created.

3. Similarly, if a dividing cycle pinches, one obtains two disconnected surfaces of genus g_i with $s_i + 1$ marked points ($i = 1, 2; g_1 + g_2 = g, s_1 + s_2 = s$). This type of degeneration also has the same number of parameters $\sum (3g_i - 3) + \sum (s_i + 1) = 3g - 4 + s$ (see Figure 1c).

It is known that $\bar{\mathcal{M}}_{g,s}$ is a compact and smooth orbifold space, and observables of topological gravity are given by the cohomology classes on $\bar{\mathcal{M}}_{g,s}$. There exist special cohomology classes introduced by Mumford and Morita, which are defined as follows: There are natural line bundles $\mathcal{L}_1, \dots, \mathcal{L}_s$ on the moduli space $\bar{\mathcal{M}}_{g,s}$. The fiber of the bundle \mathcal{L}_i at a point $\Sigma \in \bar{\mathcal{M}}_{g,s}$ is the cotangent space $T^*_{x_i} \Sigma$ to the point x_i on the surface Σ . These line bundles have the first Chern classes $c_1(\mathcal{L}_i)$ and by taking their exterior power we can define $2n$ -dimensional classes

$$\sigma_n(i) = c_1(\mathcal{L}_i)^n \in H^{2n}(\bar{\mathcal{M}}_{g,s}) \tag{7}$$

Correlation functions are defined by integrating these classes over the moduli space:

$$\langle \sigma_{n_1} \cdots \sigma_{n_s} \rangle_g \equiv \int_{\bar{\mathcal{M}}_{g,s}} c_1(\mathcal{L}_1)^{n_1} \wedge \cdots \wedge c_1(\mathcal{L}_s)^{n_s} \tag{8}$$

These integrals are topological invariants of $\bar{\mathcal{M}}_{g,s}$ and are nonzero only when the degree of the cohomology classes adds up to the dimension of the moduli space

$$\sum_{i=1}^s n_i = 3g - 3 + s \tag{9}$$

$\sigma_n(i)$ is known as the n th descendant of the puncture operator $\sigma_0(i)$, since it is associated with the marked point x_i .

The above correlation functions are evaluated using various recursion relations. First, one has the puncture equation

$$\langle \sigma_0 \sigma_{n_1} \cdots \sigma_{n_s} \rangle_g = \sum_{i=1, n_i \neq 0}^s \langle \sigma_{n_1} \cdots \sigma_{n_{i-1}} \cdots \sigma_{n_s} \rangle_g \tag{10}$$

which can be derived by considering a map $\pi: \bar{\mathcal{M}}_{g,s+1} \rightarrow \bar{\mathcal{M}}_{g,s}$ where one forgets the position of an extra point. Contributions arise when the forgotten point coincides with the other points. This relation can be used to eliminate σ_0 's from correlation functions when they are well defined. At $g=0$, less than three insertions are ill-defined and one has

$$\langle \sigma_0 \sigma_0 \sigma_0 \rangle_0 = 1 \tag{11}$$

Another basic relation is the dilaton equation for the operator σ_1 :

$$\langle \sigma_1 \sigma_{n_1} \cdots \sigma_{n_s} \rangle_g = (2g - 2 + s) \langle \sigma_{n_1} \cdots \sigma_{n_s} \rangle_g \tag{12}$$

The dilaton equation follows from the fact that since σ_1 is the first Chern class $c_1(\mathcal{L})$, it calculates the degree of the canonical line bundle of genus g surface with s punctures. At $g=1$, one insertion is required and one has

$$\langle \sigma_1 \rangle_1 = \frac{1}{24} \tag{13}$$

By combining these recursion relations, one can evaluate the correlation functions. For instance, at $g=0$ one finds

$$\langle \sigma_{n_1} \cdots \sigma_{n_s} \rangle_0 = \frac{(n_1 + \cdots + n_s)!}{n_1! \cdots n_s!} \tag{14}$$

A powerful way of computing correlation functions is given by the KdV hierarchies and Virasoro conditions as discussed below. In the context of integrable systems, it is convenient to redefine the observables as

$$\mathcal{O}_{2n+1} = (2n + 1)!! \cdot \sigma_n, \quad n \geq 0 \tag{15}$$

Topological Minimal Models

Standard intersection theory applies to the case of pure topological gravity, $p=2$. At higher values of p , the theory is generalized as follows: one introduces the coupling of topological gravity to the topological matter sector which is obtained by twisting the $\mathcal{N}=2$ superconformal theories.

We recall that $\mathcal{N}=2$ superconformal symmetry is generated by the operators, stress tensor $T(z)$, $U(1)$ current $J(z)$, and two types of supersymmetry generators $G(z)^\pm$. (In the holomorphic sector of the theory these operators depend on the holomorphic coordinate z of the Riemann surface. In the antiholomorphic sector they depend on the antiholomorphic variable \bar{z} .) Mode expansion of the stress tensor and $U(1)$ current is given by

$$T(z) = \sum_n L_n z^{-n-2}, \quad J(z) = \sum_n J_n z^{-n-1} \quad [16]$$

L_n generates the Virasoro algebra

$$[L_m, L_n] = (m-n)L_{m+n} + \frac{c}{12}m(m^2-1)\delta_{m+n,0} \quad [17]$$

where c denotes the central charge of the theory. Commutators of J_n and L_n are given by

$$[J_m, J_n] = \frac{c}{3}m\delta_{m+n,0}, \quad [L_m, J_n] = -nJ_{m+n} \quad [18]$$

It is known that there is a continuum of unitary $\mathcal{N}=2$ conformal theories in the range $c \geq 3$; however, only discrete values of the central charge $c=3k/(k+2)$, $k=1,2,\dots$ are allowed in the region $3 \geq c \geq 1$. These are the $\mathcal{N}=2$ minimal models labeled by the level k . Only a finite number of primary fields exist in these theories.

In $\mathcal{N}=2$ theory, primary fields ϕ_α are characterized by their conformal dimension and $U(1)$ charge:

$$L_0|\phi_\alpha\rangle = h|\phi_\alpha\rangle, \quad J_0|\phi_\alpha\rangle = q|\phi_\alpha\rangle \quad [19]$$

There exists a special set of primary operators, chiral primary fields ϕ_ℓ ($\ell=0,\dots,k$), which are annihilated by the supercharge operator G^+ :

$$\oint dz G^+(z)|\phi_\ell\rangle = 0 \quad [20]$$

ϕ_ℓ has the dimension and $U(1)$ charge

$$q(\phi_\ell) = \frac{\ell}{k+2}, \quad h(\phi_\ell) = \frac{1}{2}q(\phi_\ell), \quad \ell=0,1,\dots,k \quad [21]$$

By considering primary fields annihilated by G^- , we can also define antichiral fields. Antichiral fields have $U(1)$ charge opposite to those of chiral fields.

If one defines the twisted stress tensor by

$$T'(z) = T(z) + \frac{1}{2}\partial J(z) \quad [22]$$

then $T'(z)$ has a vanishing central charge. Furthermore, the conformal dimensions of the supersymmetry operators G^\pm become shifted from $3/2$ to $h(G^+)=1$ and $h(G^-)=2$. It is then possible to integrate G^+ on the Riemann surface and define a fermionic scalar operator $G_0^+ = \oint dz G^+(z)$. From the $\mathcal{N}=2$ algebra, one has

$$(G_0^+)^2 = 0, \quad \{G_0^+, G^-(z)\} = 2T'(z) \quad [23]$$

If we identify G_0^+ as the Becchi–Rouet–Stora–Tyutin (BRST) operator of the theory, then the twisted stress tensor becomes BRST trivial, which is the characteristic feature of topological field theory. Thus, we obtain a topological field theory by twisting $\mathcal{N}=2$ conformal theory (Eguchi and Yang 1990). These are topological minimal models. BRST-invariant observables are given by the chiral primary fields [20]. (To be precise, when we take account of the antiholomorphic sector, we may define either $Q = G_0^+ + \bar{G}_0^+$ or $Q = G_0^+ + \bar{G}_0^-$ as the BRST operator. Thus, in general, we obtain two different topological field theories. This is the origin of the mirror symmetry. In the context of topological gravity, one takes the convention $Q = G_0^+ + \bar{G}_0^-$.)

Now, we consider the coupling of topological gravity to topological minimal models. We identify $k=p-2$. Making use of chiral fields ϕ_ℓ ($\ell=0,\dots,p-2$), observables are constructed:

$$\sigma_{n,\ell} = \sigma_n \otimes \phi_\ell \quad [24]$$

$\mathcal{N}=2U(1)$ charge is identified as the degree of differential form of the moduli space. Thus, the degree of $\sigma_{n,\ell}$ is $n+\ell/p$. Correlation functions $\langle \prod_{i=1}^s \sigma_{n_i,\ell_i} \rangle_g$ are nonzero if the selection rule

$$\sum_{i=1}^s \left(n_i + \frac{\ell_i}{p} \right) = \left(3 - \frac{p-2}{p} \right) (g-1) + s \quad [25]$$

is obeyed.

We may assemble $\sigma_{n,\ell}$ into operators with one index \mathcal{O}_m as

$$\mathcal{O}_{np+\ell+1} = \prod_{r=0}^n (rp + \ell + 1) \cdot \sigma_{n,\ell} \quad [26]$$

where one introduces a convenient normalization factor. Note that the operators \mathcal{O}_m do not exist when $m \equiv 0 \pmod{p}$ and the corresponding parameters t_m are absent. This is a characteristic feature of p reduced KP hierarchy.

The puncture and dilaton equations for $(p, 1)$ theories read

$$\begin{aligned} & \langle \sigma_{0,0} \sigma_{n_1, k_1} \cdots \sigma_{n_s, k_s} \rangle_g \\ &= \sum_{i=1, n_i \neq 0}^s \langle \sigma_{n_1, k_1} \cdots \sigma_{n_{i-1}, k_{i-1}} \cdots \sigma_{n_s, k_s} \rangle_g \end{aligned} \quad [27]$$

$$\begin{aligned} & \langle \sigma_{1,0} \sigma_{n_1, k_1} \cdots \sigma_{n_s, k_s} \rangle_g \\ &= (2g - 2 + s) \langle \sigma_{n_1, k_1} \cdots \sigma_{n_s, k_s} \rangle_g \end{aligned} \quad [28]$$

The special terms at $g=0$ and $g=1$ are given by

$$\langle \sigma_{0,0} \sigma_{0,i} \sigma_{0,p-i-2} \rangle_0 = 1, \quad \langle \sigma_{1,0} \rangle_1 = \frac{p-1}{24} \quad [29]$$

Integrable Hierarchy

We now summarize some basic facts about the integrable hierarchy (see for instance eqn [5]). We introduce a p th order differential operator:

$$L = D^p + \sum_{i=0}^{p-2} u_i(x) D^i, \quad D \equiv \frac{\partial}{\partial x} \quad [30]$$

where the coefficient functions u_i are arbitrary functions of x . This Lax operator describes the p th generalized KdV hierarchy. We consider the time evolution of the operator L by an infinite set of commuting Hamiltonians:

$$\frac{\partial L}{\partial t_n} = [H_n, L], \quad n = 1, 2, \dots \quad [31]$$

where H_n is given by

$$H_n = \left(L^{n/p} \right)_+ \quad [32]$$

Here “+” denotes the non-negative part of a pseudodifferential operator and is defined as

$$A = \sum_{i=-\infty}^n f_i(x) D^i, \quad A_+ = \sum_{i=0}^n f_i(x) D^i \quad [33]$$

We also use the notation

$$\text{res } A = f_{-1}(x), \quad A_- = \sum_{i=-\infty}^{-1} f_i(x) D^i \quad [34]$$

Note that x is identified as the first time variable t_1 , that is, $x = t_1$.

It is a basic result of the calculus of pseudodifferential operators that the above Hamiltonians satisfy the zero-curvature condition

$$\frac{\partial H_m}{\partial t_n} - \frac{\partial H_n}{\partial t_m} + [H_m, H_n] = 0 \quad [35]$$

Note that when m is a multiple of p , H_m becomes a power of L and trivially commutes with L . Thus, the time variables t_m are absent for $n \equiv 0 \pmod{p}$. In the simple case of $p=2$, one has

$$L = D^2 + u(x) \quad [36]$$

and $H_3 = D^3 + (3/2)uD + (3/4)u'$. One finds

$$\frac{\partial L}{\partial t_3} = \frac{\partial u}{\partial t_3} = [H_3, L] = \frac{3}{2}u \frac{\partial u}{\partial x} + \frac{1}{4} \frac{\partial^3 u}{\partial x^3} \quad [37]$$

which is the standard KdV equation.

In the case of KP hierarchy, one starts with a pseudodifferential operator

$$Q = D + \sum_{i=1}^{\infty} a_i D^{-i} \quad [38]$$

and considers the time evolution equations

$$\frac{\partial Q}{\partial t_n} = [H_n, Q], \quad H_n = (Q^n)_+ \quad [39]$$

p -reduced KP hierarchy is obtained if one has

$$Q^p_- = 0 \quad [40]$$

By introducing a pseudodifferential operator K , one may bring Q to the simple derivative operator D as

$$Q = KDK^{-1} \quad [41]$$

K has an expansion of the form

$$K = 1 + \sum_{i=1}^{\infty} a_i D^{-i} \quad [42]$$

After time evolution, the coefficient functions $u_i(x)$ of the Lax operator depend also on the variables t_2, t_3, \dots and become functions of $t \equiv \{t_1, t_2, \dots\}$. These functions are expressed by the τ -function $\tau(t)$ of the hierarchy in the following manner:

$$\text{res } K = -\frac{\partial}{\partial x} \log \tau(t) \quad [43]$$

$$\text{res } L^{i/p} = \frac{\partial^2}{\partial x \partial t_i} \log \tau(t) \quad [44]$$

These residues are expressed in terms of $\{u_i\}$ and their derivatives in x , and one can determine them in terms of the τ -function.

In the case $p=2$, one has

$$[H_k, L] = 2D \operatorname{res}(L^{k/2}) = DR_k, \quad k = \text{odd} \quad [45]$$

Here $\{R_k\}$ are the Gelfand–Dikii potentials

$$\begin{aligned} R_1 &= u, & R_3 &= \frac{1}{4}(3u^2 + u'') \\ R_5 &= \frac{1}{16}(10u^3 + 5u'^2 + 10uu'' + u''') \\ &\vdots \end{aligned} \quad [46]$$

and obey the recursion relation

$$DR_{k+2} = \frac{1}{4}(D^3 + 2(Du + uD))R_k \quad [47]$$

If one uses the relation [44], Gelfand–Dikii potentials are identified as

$$R_k = 2\langle \mathcal{O}_1 \mathcal{O}_k \rangle \quad [48]$$

By setting $k=1$, we note $u = 2\langle \mathcal{O}_1 \mathcal{O}_1 \rangle$ and find that the evolution equations [31] are all satisfied as

$$\begin{aligned} \frac{\partial L}{\partial t_k} &= \frac{\partial u}{\partial t_k} = 2 \frac{\partial}{\partial t_k} \langle \mathcal{O}_1 \mathcal{O}_1 \rangle = 2D \langle \mathcal{O}_1 \mathcal{O}_k \rangle \\ &= DR_k = [H_k, L] \end{aligned} \quad [49]$$

Now it is possible to identify the initial condition for the Lax operator in the case of topological $(p, 1)$ gravity. By using the definition

$$\log \tau(t) = \sum_{g=0}^{\infty} \left\langle \exp \sum_n t_n \mathcal{O}_n \right\rangle_g \quad [50]$$

one has

$$\operatorname{res} L^{i/p}(0) = \langle \mathcal{O}_1 \mathcal{O}_i \rangle, \quad i = 1, \dots, p-1 \quad [51]$$

From [29] one finds

$$\operatorname{res} L^{i/p}(0) = ix \cdot \delta_{i,p-1} \quad [52]$$

This gives the initial value of the Lax operator:

$$L(0) = D^p + px \quad [53]$$

Thus, only the lowest term $u_0(x) = px$ is nonzero and higher coefficients all vanish at $t=0$. This is the special simplification which takes place in the topological gravity theory.

We note a relation

$$\left[\frac{1}{p} D, L(0) \right] = 1 \quad [54]$$

This is the so-called “string equation” (at $t=0$). At nonzero values of t , the string equation takes the form

$$\begin{aligned} [P, L] &= 1 \\ P &= \frac{1}{p} \left((L^{1/p})_+ - \sum_{k=p+1}^{\infty} kt_k (L^{(k-p)/p})_+ \right) \end{aligned} \quad [55]$$

From [55], we see that $(p, 1)$ theory corresponds to the background value of the coupling $t_{p+1} = -1/(p+1)$. In the case of (p, q) theory, background value is given by $t_{pq+1} = -1/(pq+1)$.

Virasoro Conditions

A powerful algebraic machinery controlling the structure of 2D gravity is the so-called “Virasoro conditions.” One introduces differential operators

$$L_{-1} = -\frac{\partial}{\partial t_1} + \sum_{k=p+1}^{\infty} kt_k \frac{\partial}{\partial t_{k-p}} + \frac{1}{2} \sum_{i+j=p} ij t_i t_j \quad [56]$$

$$L_0 = -\frac{\partial}{\partial t_{p+1}} + \sum_{k=1}^{\infty} kt_k \frac{\partial}{\partial t_k} + \frac{p^2 - 1}{24} \quad [57]$$

By using the fact that derivative in t_n brings down the operator \mathcal{O}_n when acting on the τ -function, it is easy to show that

$$L_{-1} \cdot \tau = 0 \quad [58]$$

$$L_0 \cdot \tau = 0 \quad [59]$$

reproduce the puncture [27] and dilaton equation [28], respectively. It is possible to show that the L_{-1} -condition, $L_{-1} \cdot \tau = 0$, is equivalent to the string equation [55].

Together with the operators ($n \geq 1$)

$$L_n = -\frac{\partial}{\partial t_{1+(n+1)p}} + \sum_{k=1}^{\infty} kt_k \frac{\partial}{\partial t_{k+np}} + \frac{1}{2} \sum_{i+j=np} \frac{\partial^2}{\partial t_i \partial t_j}$$

they generate Virasoro algebra ($L'_n \equiv (1/p)L_n$)

$$[L'_m, L'_n] = (m-n)L'_{m+n}, \quad n, m \geq -1 \quad [60]$$

It is possible to show that the $(p, 1)$ model obeys the Virasoro conditions [6]

$$L_n \cdot \tau = 0, \quad n \geq -1 \quad [61]$$

It is known that $(p, 1)$ models with $p > 2$ also obey constraints of W -algebra.

The relationship of the Virasoro conditions to KdV hierarchy is summarized as

$$\begin{aligned} &\text{string equation} + \text{KdV hierarchy} \\ &\iff \text{Virasoro and } W\text{-algebra constraints} \end{aligned}$$

Topological σ -Model

It is known that when the target space of a supersymmetric nonlinear σ -model is a Kahler manifold K , the theory acquires an enhanced $\mathcal{N}=2$

supersymmetry. Then we can twist the theory and converted into a topological field theory. This is the topological σ -model [7]. The partition function of the theory consists of a sum over world-sheet instantons, that is, holomorphic maps from the Riemann surface to the target space K . Due to supersymmetry, functional determinants around instantons cancel and the theory simply counts the number of holomorphic curves inside the Kahler manifold K . Thus, the topological σ -model has a close relationship with enumerative problems in algebraic geometry, that is, Gromov–Witten invariants and quantum cohomology theory.

When the topological σ -model is coupled to topological gravity, the BRST-invariant observables are given by $\sigma_n(\Phi_i) \equiv \sigma_n \otimes \Phi_i$, where Φ_i are cohomology classes of K . Correlation functions are defined as

$$\left\langle \prod_{i=1}^s \sigma_{n_i}(\Phi_i) \right\rangle_{g,d} = \int_{\bar{\mathcal{M}}(K;d)_{g,s}} \prod_{i=1}^s c_i(\mathcal{L}_i)^{n_i} \wedge e_i^*(\Phi_i) \quad [62]$$

Here $\bar{\mathcal{M}}_{g,s}(K;d)$ denotes the (stable compactification of) moduli space of degree d holomorphic maps to K from genus g Riemann surfaces Σ . e_i^* is the pullback of the evaluation map $e_i: (f; x_1, \dots, x_s) \in \mathcal{M}_{g,s}(K;d) \rightarrow f(x_i) \in K$ by f where f is a holomorphic map. Correlation functions [62] give topological (symplectic) invariants of the manifold K . In the cases $n_i = 0$ ($i = 1, \dots, s$), they are known as Gromov–Witten invariants.

Equation [62] is nonvanishing if the selection rule

$$\sum_{i=1}^s (n_i + q_i) = \dim \mathcal{M}_{g,s}(K;d) = c_1(K)d + (3 - \dim K)(g - 1) + s \quad [63]$$

is obeyed, where q_i is the degree of cohomology class Φ_i and $c_1(K)$ is the first Chern class of the tangent bundle of K .

We see that there is a close parallel between the topological σ -model and $(p, 1)$ topological gravity. If we formally set $q_i = \ell_i/p$, $c_1(K) = 0$, and $\dim K = (p - 2)/p$, eqn [63] agrees with eqn [25]. Based on this analogy, Eguchi, Hori, and Xiong proposed the Virasoro conjecture [8], that is, generating functions of the number of holomorphic maps to arbitrary

Kahler manifolds are annihilated by the Virasoro operators which are constructed by taking an analogy with those of $(p, 1)$ gravity. The Virasoro conjecture is a natural generalization of Witten’s conjecture, and has recently been rigorously proved in the case of curves and projective spaces.

Excellent reviews on the theory of 2D topological gravity are given in Witten (1991) and Dijkgraaf (1991).

See also: Axiomatic Approach to Topological Quantum Field Theory; Large- N and Topological Strings; Mirror Symmetry: A Geometric Survey; Moduli Spaces: An Introduction; Riemann Surfaces; Topological Sigma Models; WDVV Equations and Frobenius Manifolds.

Further Reading

Brézin E and Kazakov V (1990) Exactly soluble field theory of closed strings. *Physics Letters B* 236: 144.

Dijkgraaf R (1991) *Lectures Presented at Cargèse Summer School on New Symmetry Principles in Quantum Field Theory*, hep-th/9201003.

Dijkgraaf R, Verlinde E, and Verlinde H (1991) Loop equations and Virasoro constraints in non-perturbative 2d quantum gravity. *Nuclear Physics B* 348: 435.

Douglas M and Shenker S (1990) Strings in less than one dimension. *Nuclear Physics B* 335: 635.

Eguchi T and Yang S-K (1990) $\mathcal{N}=2$ superconformal models as topological field theories. *Modern Physics Letters A* 5: 1693.

Eguchi T, Hori K, and Xiong C-S (1997) Quantum cohomology and Virasoro algebra. *Physics Letters B* 402: 71.

Fukuma M, Kawai H, and Nakayama R (1991) Continuum Schwinger–Dyson equations and universal structures in two-dimensional quantum gravity. *International Journal of Modern Physics A* 6: 1385.

Gelfand IM and Dikii LA (1975) Asymptotic behavior of the resolvent of Sturm–Liouville equations and the algebra of the Korteweg–de Vries equations. *Russian Mathematical Surveys* 30(5): 77–113.

Gross D and Migdal A (1990) Non-perturbative two dimensional quantum gravity. *Physical Review Letters* 64: 17.

Kontsevich M (1991) Intersection theory on the moduli space of curves and the matrix Airy function. *Communications in Mathematical Physics*.

Witten E (1988) Topological sigma models. *Communications in Mathematical Physics* 118: 411.

Witten E (1990) On the topological phase of two dimensional gravity. *Nuclear Physics B* 340: 281.

Witten E (1991) Two-dimensional gravity and intersection theory of moduli space. *Surveys in Differential Geometry* 1: 243–310.

Topological Knot Theory and Macroscopic Physics

L Boi, EHESS and LUTH, Paris, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction to the Physical and Mathematical Contexts and Issues

One of the most exciting developments of mathematical physics in the last three decades has been the discovery of numerous intimate relationships between the topology and the geometry of knot theory and the dynamics of many domains of “classical” and “new” macroscopic physics. Indeed, complex systems of knotted and entangled filamentary structures are ubiquitous in nature and arise in such disparate contexts as electrodynamics, magnetohydrodynamics, fluid dynamics (vortex structures), superfluidity, dynamical systems, plasma physics, cosmic string theory, chaos of magnetic flows and nonlinear phenomena, turbulence, polymer physics, and molecular biology. In the recent years, mathematical tools have been developed to identify and analyze the geometrical and topological complex structures and behaviors of such systems and relate this information to energy levels and stable states.

The influence of geometry and topology on macroscopic physics has been especially fruitful in the study and comprehension of the following topics.

1. *Knots and braids in dynamical systems.* It is now clear that the chaotic behavior of the Hénon–Heiles system and other nonlinear systems is driven and controlled by topological properties. For example, it has been found that trajectories in the phase space form hyperbolic knots. The finding of knots in the Lorenz equations is another important theme closely related to the previous. By varying the Rayleigh number r , a parameter in the Lorenz equations, both chaotic and periodic behavior is observed. In the recent years, the knots (notably several torus knots) corresponding to the different periodic solutions of the system have been found and classified. By finding hyperbolic knots and in particular hyperbolic figure-8 knot as a solution to the Lorenz equations the suspicion that there exists a new route to chaos would be strengthened.

2. *Topological structures of electromagnetic fields.* Progress in the field of space physics, astronomy, and astrophysics over the last decade, increasingly reveals the significance of topological magnetic fields in these areas. In particular, the interaction of plasma and magnetic field can create an astonishing variety of structures, which often exhibit linked and knotted

forms of magnetic flux. In these complex structures of the fields, huge amounts of magnetic energy can be stored. It is, however, a typical property of astrophysical plasmas, that the dynamics of magnetic fields is alternating between an ideal motion, where all forms of knottedness and linkage of the field are conserved (topology conservation), and a kind of disruption of the magnetic structure, the so-called magnetic reconnection. In the latter, the magnetic structure breaks up and reconnects, a process often accompanied by explosive eruptions, where enormous amounts of energy are set free. Magnetic reconnection is in close analogy to splitting of knots, which makes us confident that the global dynamics of magnetic and electromagnetic fields can be characterized with the help of such topological quantities as well.

3. *Knotting and unknotting of phase singularities.* It has long been known that dislocation lines can be closed, and recently it was shown that they can be knotted and linked. Moreover, Berry and Dennis (2001) constructed exact solutions of the Helmholtz equation representing torus knots and links; in fact, a straightforward application of this idea led to knotted and linked dislocation lines in stationary states of electrons in hydrogen. As a parameter, called α , is varied, the topology of dislocation lines can change, leading to the creation of knots and links from initially simple dislocation loops, and the reverse process of unknotting and unlinking. The main purpose here is to elucidate the mechanism of these changes of topology. All waves are solutions of monochromatic wave equations, that is, stationary waves, and α is an external parameter that could be manipulated experimentally. However, α could represent time, and then the analogous solutions of time-dependent wave equations would describe knotting and linking events in the history of waves. The methods of Berry and Dennis are based on exact stationary solutions of wave equations, and lead to knots and links threaded by multistranded helices.

The Origins of Topological Vortex Dynamics Ideas

The intimate relationship between three-dimensional vortex dynamics and topology was recognized as early as 1869 by W Thomson (Lord Kelvin) who tried to elaborate a theory of matter in which atoms were thought to be tiny vortex filaments embedded in an elastic-like fluid medium, called ether. Accordingly, the infinite variety of possible chemical compounds was given by the endless family of topological

combinations of linked and knotted vortices. Kelvin was inspired by the work of Gauss, who in an attempt to describe topologically the behavior of two inseparably closed linked circuits carrying electric current, found a relationship between the magnetic action induced by the currents and a pure number that depends only on the type of link, and not on the geometry: this number is the first topological invariant now known as the linking number.

In modern mathematical terms, Gauss introduced an invariant of a link consisting of two simple closed curves γ_1, γ_2 in R^3 , namely the signed number of turns of one of the curves around the other, the linking coefficient $\{\gamma_1, \gamma_2\}$ of the link. His formula for this is

$$\begin{aligned} N &= \{\gamma_1, \gamma_2\} \\ &= \frac{1}{4\pi} \int_{\gamma_1} \int_{\gamma_2} ([d\gamma_1(t), d\gamma_2(t)], \gamma_1 - \gamma_2) / \\ &\quad |\gamma_1(t) - \gamma_2(t)|^3 \end{aligned} \quad [1]$$

where $[,]$ denotes the vector (or cross) product of vectors in R^3 and $(,)$ the Euclidean scalar product. Thus, this integral always has an integer value N . If we take one of the curves to be the z -axis in R^3 and the other to lie in the (x, y) -plane, then the formula [1] gives the net number of turns of the plane curve around the z -axis. It is interesting to note that the linking coefficient [1] may be zero even though the curves are nontrivially linked. Thus, its having nonzero value represents only a sufficient condition for nontrivial linkage of the loops. This last consideration leads naturally to the mathematical concepts of knots and links whose most striking properties have been investigated in our introductory article (see Mathematical Knot Theory).

The other source of inspiration of Kelvin's theory of matter was the Helmholtz's laws of vortex motion, which state that in an ideal fluid (where there is no viscosity) vortex lives forever: two closed vortex rings, once linked, will always be linked. The classical results obtained by Helmholtz are basic to understanding the dynamics of Euler motions. The vorticity of a velocity field is its curl and is denoted $\omega_t(z) := \text{curl}(X(z, t))$. In two dimensions, the vorticity is a real-valued function and $\omega_t = -\Delta\Psi$, where Ψ is the stream function of $X(z, t)$. Recall that the push-forward of a scalar field (0-form) s under a diffeomorphism f is $f_*s = s \circ f^{-1}$. These results, in modern terms, can be stated as follows:

Theorem (Helmholtz–Kelvin). *An incompressible fluid motion (M_t, ϕ_t) with velocity field X and vorticity ω_t is Euler if and only if its vorticity is passively transported,*

$$\phi_t^* \omega_0 = \omega_t$$

and circulation around all smooth simple closed curves C are preserved under the flow,

$$\frac{d}{dt} \int_{\phi_t(C)} X \cdot d\mathbf{r} = 0$$

One knows that in three dimensions, the Helmholtz–Kelvin theorem says that the vorticity (now a vector field) is transported. Thus, with generic initial vorticity a 3D time-periodic Euler fluid motion preserves a nontrivial vector field. One very interesting question that remains to be elucidated is the following: are there any chaotic, time-periodic Euler flows with stationary boundaries?

The Connection between Topological and Numerical Invariants of Knots and the Physical Helicity of Vector Fields

The writhing number of a curve in Euclidean three-dimensional space is the standard measure of the extent to which the curve wraps and coils around itself; it has proved its importance for electrodynamics and fluid mechanics in the study of the knotted structures of magnetic vortices and dynamics flows, and for molecular biologists in the study of knotted duplex DNA and the enzymes which affect it. The helicity of a divergenceless vector field defined on a domain in Euclidean 3-space, introduced by Woltjer in 1958 in an astrophysical context and coined by Moffat in 1969 in the study of its topological meaning, is the standard measure of the extent to which the field lines wrap and coil around one another; it plays important roles in fluid mechanics, magnetohydrodynamics, and plasma physics. The “Biot–Savart operator” associates with each current distribution on a given domain the restriction of its magnetic field to the domain. When the domain is simply connected, the divergence-free fields which are tangent to the boundary and which minimize energy for given helicity provide models for stable force-free magnetic fields in space and laboratory plasmas; these fields appear mathematically as the extreme eigenfields for an appropriate modification of the Biot–Savart operator. Information about these fields can be converted into bounds on the writhing number of a given piece of DNA.

Recent researches (Cantarella *et al.* 2001) obtained rough upper bounds for the writhing number of a knot or link in terms of its length and thickness, and rough upper bounds for the helicity of a vector field in terms of its energy and the geometry of its domain. It was also showed that in the case of classical electrodynamics in vacuum, the

natural helicity invariant, called the electromagnetic helicity, has an important particle meaning: the difference between the numbers of right- and left-handed photons. Recently, a topological model of classical electrodynamics has been proposed in which the helicity is topologically quantized, in a relation that connects the wave and particle aspects of the fields (Trueba and Rañada 2000).

Consider two disjoint closed space curves, C and C' , and the Gauss' integral formula for their linking number

$$\text{Lk}(C, C') = \frac{1}{4\pi} \int_{C \times C'} \frac{(dx/ds \times dy/dt) \times x - y/|x - y| ds dt}{|x - y|^3} \quad [2]$$

The curves C and C' are assumed to be smooth and to be parametrized by arclength. Now the question is to know what happens to this integral when the two space curves C and C' come together and coalesce as one curve C . At first glance, the integrand looks like it might blow up along the diagonal of $C \times C'$, but a careful calculation shows that in fact the integrand approaches zero on the diagonal, and so the integral converges. Its value is the writhing number $\text{Wr}(C)$ of C defined above:

$$\text{Wr}(C) = \frac{1}{4\pi} \int_{C \times C} \frac{(dx/ds \times dy/dt) \times x - y/|x - y| ds dt}{|x - y|^3} \quad [3]$$

Here is the very useful result, due to Fuller (1978). The writhing number of a knot K is the average linking number of K with its slight perturbations in every possible direction:

$$\text{Wr}(K) = \frac{1}{4\pi} \int_{W \in S^2} \text{Lk}(K, K + \varepsilon W) d(\text{area}) \quad [4]$$

This is helpful for getting a quick approximation to the writhing number of a knot which almost lies in a plane; in the example of a trefoil knot, $\text{Wr}(K) \approx 3$.

Here, a very important result must be recalled, a "bridge theorem," proved by Berger and Field (1984), see also Ricca and Moffatt (1992), which connects helicity of vector fields to writhing of knots and links, and which can be used to convert upper bounds on helicity into upper bounds on writhing.

Proposition (Berger and Field). *Let K be a smooth knot or link in 3-space and $\Omega = N(K, R)$ a tubular neighborhood of radius R about K . Let V be a vector field defined in Ω , orthogonal to the cross-sectional disks, with length depending only on distance from K . This makes V divergence-free and tangent to the boundary of Ω . Then the writhing number $\text{Wr}(K)$ of K and the helicity $H(V)$ of the vector field V are related by the formula*

$$H(V) = \text{Flux}(V)^2 \text{Wr}(K)$$

In the formula, $\text{Flux}(V)$ denotes the flux of V through any of the cross-sectional disks D ,

$$\text{Flux}(V) = \int_D V \cdot n d(\text{area})$$

where n is a unit normal vector field to D .

A key feature of this formula is that the helicity of V depends on the writhing number of K , but not any further on its geometry; in particular, such quantities as the curvature and torsion of K do not enter into the formula. Berger and Field actually showed that the helicity $H(V)$ is a sum of two terms: a "kink helicity," which is given by the right-hand side of the above formula, and a "twist helicity," which is easily shown in our case to be zero. Their proof assumes K is a knot, but it is straightforward to extend it to cover links.

Let Ω be a compact domain in 3-space with smooth boundary $\partial\Omega$; we allow both Ω and $\partial\Omega$ to be disconnected. Let V be a smooth vector field (where "smooth" means of class C^∞), defined on the domain Ω . The helicity $H(V)$ of the vector field V is defined by the formula

$$H(V) = \frac{1}{4\pi} \int_{\Omega \times \Omega} \frac{V(x) \times V(y) \cdot x - y/|x - y|^3}{|x - y|^3} d(\text{vol})_x d(\text{vol})_y \quad [5]$$

Clearly, helicity for vector fields is the analog of writhing number for knots. Both formulas are variants of Gauss' integral formula for the linking number of two disjoint closed space curves.

In order to understand this formula for helicity, think of V as a distribution of electric current, and use the Biot-Savart law of electrodynamics to compute its magnetic field:

$$\text{BS}(V)(y) = \frac{1}{4\pi} \int_{\Omega} \frac{V(x) \times y - x/|y - x|^3}{|y - x|^3} d(\text{vol})_x \quad [6]$$

Then the helicity of V can be expressed as an integrated dot product of V with its magnetic field $\text{BS}(V)$:

$$\begin{aligned} H(V) &= \frac{1}{4\pi} \int_{\Omega \times \Omega} \frac{V(x) \times V(y) \cdot x - y/|y - x|^3}{|y - x|^3} d(\text{vol})_x d(\text{vol})_y \\ &= \int_{\Omega} V(y) \cdot \left[\frac{1}{4\pi} \int_{\Omega} \frac{V(x) \times y - x/|x - y|^3}{|x - y|^3} d(\text{vol})_x \right] d(\text{vol})_y \\ &= \int_{\Omega} V(y) \cdot \text{BS}(V)(y) d(\text{vol})_y \\ &= \int_{\Omega} V \cdot \text{BS}(V) d(\text{vol}) \end{aligned}$$

Cantarella *et al.* (2001) found two very interesting results.

Theorem 1 *Let K be a smooth knot or link in 3-space, with length L and with an embedded tubular neighborhood of radius R . Then the writhing number $Wr(K)$ of K is bounded by*

$$|Wr(K)| < 1/4(L/R)^{4/3}$$

For the proof, see Cantarella *et al.* (2001).

Theorem 2 *The helicity of a unit vector field V defined on the compact domain Ω is bounded by*

$$|H(V)| < 1/2 \text{vol}(\Omega)^{4/3}$$

Let us now give a brief overview of the methods used to find sharp upper bounds for the helicity of vector fields defined on a given domain Ω in 3-space. As usual, Ω will denote a compact domain with smooth boundary in 3-space. Let $K(\Omega)$ denotes the set of all smooth divergence-free vector fields defined on Ω and tangent to its boundary. These vector fields, sometimes called “fluid knots,” are prominent for several reasons: (1) They are natural vector fields to study in a “fluid dynamics approach” to geometric knot theory. (2) They correspond to incompressible fluid flows inside a fixed container. (3) They are vector fields most often studied in plasma physics. (4) For given energy (equivalently minimize energy for given helicity), they provide models for stable force-free magnetic fields in gaseous nebulae and laboratory plasmas. (5) The search for these helicity-maximizing fields can be converted to the task of solving a system of partial differential equations. (6) The fluid knots can reveal some fundamental and still unknown mechanisms, which characterize the phenomenon of phase transition, and in particular the transition from chaotic (unstable) phases and behaviors of matter to ordered (stable) ones.

Knots and Fluid Mechanics (Vortex Lines, Magnetic Helicity, and Turbulence)

The Kelvin’s theory of explaining atoms as knotted vortices in fluid ether was seminal in the development of topological fluid mechanics. The recent revival (starting in the 1970s) is mainly due to the work of Moffat, on topological interpretation of helicity, and Arnol’d, on asymptotic linking number of space-filling curves. Modern developments have been influenced by recent progress in the theory of knots and links.

Influence of Geometry and Topology on Fluid Flows

Ideal topological fluid mechanics deals essentially with the study of fluid structures that are continuously deformed from one configuration to another by ambient isotopies. Since the fluid flow map φ is both continuous and invertible, then $\varphi_{t_1}(K)$ and $\varphi_{t_2}(K)$ generate isotopies of a fluid structure K (e.g., a vortex filament) for any $\{t_1, t_2\} \in I$. Isotopic flows generate equivalence classes of (linked and knotted) fluid structures. In the case of (vortex or magnetic) fluid flux tubes, fluid actions induce continuous deformations in D . One of the simplest deformations is local stretching of the tube. From a mathematical viewpoint, this deformation corresponds to a time-dependent, continuous reparametrization of the tube centerline. This reparametrization (via homotopy classes) generates ambient isotopies of the flux tube, with a continuous deformation of the integral curves.

Moreover, in the context of the Euler equations, the Reidemeister moves (or isotopic plane deformations), whose changes conserves the knot topology, are performed quite naturally by the action of local flows on flux tube strands. If the fluid in $(D - K)$ is irrotational, then these fluid flows (with velocity u) must satisfy the Dirichlet problem for the Laplacian of the stream function φ , that is,

$$\begin{aligned} u &= \nabla \varphi \quad \text{in } (D - K) \\ \nabla^2 \varphi &= 0 \end{aligned} \quad [7]$$

with normal component of the velocity to the tube boundary u_\perp given. Equations [7] admit a unique solution in terms of local flows, and these flows are interpretable in terms of Reidemeister’s moves performed on the tube strands. Note that boundary conditions prescribe only u_\perp , whereas no condition is imposed on the tangential component of the velocity. This is consistent with the fact that tangential effects do not alter the topology of the physical knot (or link). The three type of Reidemeister’s moves are therefore performed by local fluid flows, which are solutions to [7], up to arbitrary tangential actions.

Knotted and Linked Tubes of Magnetic Flux

Let T be the standard solid torus in R^3 given by

$$((2 + \varepsilon \cos \theta) \cos \varphi, (2 + \varepsilon \cos \theta) \sin \varphi, \varepsilon \sin \theta)) \quad [8]$$

where $0 \leq \theta, \varphi < 2\pi$, and $0 \leq \varepsilon < 1$. For relatively prime integers p and q , let $F_{p,q}$ denote the foliation

of T by the curves $\gamma_{\varepsilon,\theta}$ (where $0 \leq \varepsilon \leq 1$ and $0 \leq \theta < 2\pi$) given by

$$\begin{aligned} \gamma_{\varepsilon,\theta}(s) &= (2 + \varepsilon \cos(\theta + qs)) \cos(ps), \\ (2 + \varepsilon \cos(\theta + qs)) \sin(ps), &\varepsilon \sin(\theta + qs) \end{aligned} \quad [9]$$

where $0 \leq s < 2\pi$.

Definition A magnetic tubular link (or magnetic link) is a smooth immersion into \mathbb{R}^3 of finitely many disjoint standard solid tori $\cup_{i=1}^n T_i$

$$L : \cup_{i=1}^n T_i \rightarrow \mathbb{R}^3$$

and a smooth magnetic field B on \mathbb{R}^3 such that

- (i) L is an imbedding when restricted to the interior of $\cup_{i=1}^n T_i$,
- (ii) the bounding surface of $\cup_i L(T_i)$, that is, $\cup_i L(\partial T_i)$ is a magnetic surface, and
- (iii) for each component LT_i , there exist relatively prime nonzero integers p_i and q_i such that L maps the foliation F_{p_i, q_i} of T_i onto the integral curves of B in LT_i .

Remark Thus, for every fixed i and j , the linking number between an arbitrary field line in LT_i and an arbitrary field line in LT_j is the same regardless of which integral curves are chosen from LT_i and LT_j , respectively. This is true even when $i=j$.

It follows that a magnetic link $\cup_i LT_i$ remains a magnetic link under the action of the fluid flow, that is, $\cup_i g_t LT_i$ is a magnetic link for $t \geq 0$.

Keeping that the magnetic field B is frozen in the fluid, we can now find and study those properties of magnetic links that are invariant under the action of fluid flow. One obvious invariant is the volume V_i of each flux tube $g_t LT_i$, that is,

$$V_i = \text{Vol}(LT_i) = \text{Vol}(g_t LT_i) = \iint \int_{g_t LT_i} d(\text{vol}) \quad [10]$$

which remains unchanged because of incompressibility.

Another invariant of fluid flow is defined as follows:

Definition Let L be a magnetic link. For each solid torus T_i , choose a meridional disk D_i . The magnetic flux $\Phi_i = \Phi(LT_i)$ in the i th component is the surface integral defined as

$$\Phi_i = \Phi(LT_i) : \iint_{LD_i} B \cdot U d(\text{area})$$

where U denotes the normal to the surface LD_i pointing in the positive direction induced by the B field.

It can be shown that Φ_i is independent of the chosen meridional disk. It also can be shown that each Φ_i is a fluid flow invariant, that is,

$$\Phi_i(g_t LT_i) = \iint \int_{g_t LD_i} B \cdot U d(\text{area}) \quad [11]$$

is independent of t .

One more fluid invariant that will play a central role in the energy minimization of magnetic links is given by the following definition.

Definition The helicity of a magnetic link L is defined as

$$H(L) = \iint \int_{\cup_i LT_i} A \cdot B d(\text{vol})$$

The term helicity was first introduced in a fluid context by Moffat, and it was previously used in particle physics for the scalar product of the momentum and spin of a particle. In another connection, note that the helicity $H(L)$ is the same as the Chern–Simons action:

$$\begin{aligned} H(L) &= \int A \wedge dA \\ &= \int \text{tr}(A \wedge dA + \frac{2}{3} A \wedge A \wedge A) \end{aligned} \quad [12]$$

where A now denotes the magnetic vector potential as a 1-form.

It can be shown that $H(L)$ is gauge invariant, and hence well defined.

Theorem (Moffat). *The helicity is invariant under fluid flow, that is,*

$$\frac{d}{dt} H(g_t L) = 0$$

Arnol'd (1998) defines the helicity in a more abstract setting and shows that it is invariant under the group $S(\text{Diff})$ of volume-preserving diffeomorphisms.

The following theorem summarizes the many results due to Moffat, Ricca, Berger, Lomonaco, Hornig, Kauffman, and others, relating the helicity of magnetic links to linking and to magnetic flux.

Theorem *Let L be a magnetic link. Then*

$$H(L) = \sum_{1 \leq i \leq n} \Phi_i^2 \text{SL}_{F_i} + 2 \sum_{1 \leq i < j \leq n} \Phi_i \Phi_j \text{LK}_{ij}$$

where SL_{F_i} denotes the self-linking number of the axis curve of the tube LT_i with respect to the framing F_i induced by the integral curves of the magnetic field B within LT_i , and LK_{ij} denotes the linking number between any integral curve of

the magnetic field B in LT_i with any integral curve of the magnetic field B in LT_j .

Remark In fact, SL_{F_i} is the same as the linking number between any two integral curves of the magnetic field B within the tube LT_i .

Thus, as many authors have showed, the helicity does reflect the topology and the geometry of the magnetic lines of force within a magnetic link. If, for example, L has only one component, that is, L is a magnetic knot, then

$$H(L) = \Phi^2 SL_F(C) \tag{13}$$

where $SL_F(C)$ is the self-linking number of the axis curve C of the knotted tube with respect to the framing F induced by the integral curves of the magnetic field B within the magnetic knot. If, for example, the tube is knotted in the form of a trefoil and if the magnetic lines of force appear to be parallel to the axis curve when the trefoil is placed on a plane flat surface, then $SL = \pm 3$ and

$$H = \pm 3\Phi^2 \tag{14}$$

On the other hand, if for example, the magnetic lines of force induce the trivial framing in each component, then

$$H(L) = 2 \sum_{1 \leq i < j \leq n} \Phi_i \Phi_j LK_{ij} \tag{15}$$

Thus, if L is a magnetic two-component Hopf link with no twisting of the integral curves of the magnetic field within the components of L , then

$$H(L) = \pm 2\Phi_1 \Phi_2 \tag{16}$$

because the self-linking number based on the B -field framing is zero for each component, and the linking number between the two components is ± 1 .

Energy of Magnetic Knots and Links

Let us conclude this section with the definition of the energy of a magnetic link.

Definition The magnetic energy $E_M(L)$ of a magnetic link L is defined by the classical formula

$$E_M(L) = \frac{1}{8\pi} \int \int \int_{\cup_i LT_i} |B|^2 d(\text{vol}) \text{ (Gaussian units)}$$

Although the energy E_M is not flow invariant, it will play a central role in magnetic relaxation of knots and minimum energy magnetic links.

Consider a magnetic link L in a perfectly conducting, incompressible, viscous fluid. As a result of dissipative frictional fluid forces, the magnetic energy $E_M(g_t L)$ of $g_t L$ will decrease with time t . In

losing energy, the magnetic lines of force will contract. On the other hand, since this is a volume-preserving process, the cross sections of the flux tubes of $g_t L$ will at the same time expand. These changes of topology occur while the flux Φ , volume V , and helicity of $g_t L$ will remain the same. In other words, knotted magnetic flux tubes left free to evolve in such a fluid will do so by conserving their magnetic flux Φ and volume V , but converting their magnetic energy into kinetic energy, which in turn dissipates by internal friction. Magnetic links and knots evolve from high to low magnetic energy levels, conserving topology; and because of the induced shortening of field lines under conservation of volume, they become fatter and fatter, with an increase of the average tube cross section.

This process cannot continue indefinitely. Eventually, the magnetic flux tubes of $g_t L$ must make contact with each other. In other words, the topology of the magnetic link $g_t L$, as expressed in knotting and linking, creates a barrier to the full dissipation of the magnetic link's energy, that is, $E_M(g_t L)$ has a positive lower bound that results from the topology of $g_t L$. That means, in other words, that relaxation is obstructed by the knottedness and entanglement of the field lines, and a minimum magnetic energy is reached. Thus, the magnetic link will reach a nontrivial stable and invariant energy state, much as Kelvin conjectured his atomic vortices would.

Various estimates of magnetomechanical energy in terms of topological quantities have been put forward in recent years (see Freedmann and He (1994)). These relations give lower bounds for the energy levels attainable by knot or link types by taking into account the effects that linking numbers and number of crossings have on the energy of the relaxed state. These bounds are expressed by relationship of the kind

$$E_{\min} \geq \phi(C_{\min}, \Phi, V, N) \tag{17}$$

where E_{\min} is the equilibrium energy and ϕ gives the relationship between physical quantities – such as total flux Φ , number of tubes N , magnetic volume V – and topology, given here by the minimal possible number of crossing C_{\min} . These relations offer numerous advantages due to the explicit dependence on qualitative properties of the flow field. A simple example is provided by the analysis of three braids, which shows that magnetic energy grows quadratically in time due to random braiding. This means that the least possible amount of magnetic energy that can be attained by the physical knot or link is determined purely by its topology. If topological information sets the levels of minimum energy accessible to the knot or link, geometric properties

may also influence the relaxation process. Considerations of helicity and linking numbers, for example, demonstrate that internal rearrangement of magnetic field geometry leads to a spectrum of different asymptotic endstates with the same topology. Moreover, magnetic knots have a natural tendency to get rid of excessive torsion of field lines and S-shaped tube geometries, and this may influence the relaxation process.

Since the helicity $H(g_i L)$ is both an invariant of fluid flow and an expression of the magnetic link $g_i L$'s topology, the following theorem, first stated by Moffat, is a mathematical expression of this topological bounds.

Theorem *Let L be a magnetic link. Then*

$$E_M(L) \geq q_0 |H(L)|$$

where q_0 is a nonzero constant that is independent of the magnetic link.

Freedman and He (1991) obtain more subtle and tighter topological bounds on the minimum energy of magnetic links. For example, for a magnetic knot K , they prove that

$$\begin{aligned} E_M(K) &\geq \frac{1}{4\pi^{5/4}} \frac{\Phi(K)^{3/2} \text{ac}(K)^{3/4}}{V(K)^{1/3}} \\ &\geq \frac{1}{4\pi^{5/4}} \frac{\Phi(K)^{3/2} (2g(K) - 1)}{V(K)^{1/3}} \\ &\quad \text{(Gaussian units)} \end{aligned} \quad [18]$$

where $V(K)$ denotes the volume of the magnetic knot K , $\Phi(K)$ denotes the flux in K , $\text{ac}(K)$ is the asymptotic crossing number, and $g(K)$ is the genus of the knot K . Freedman and He conjecture that $\text{ac}(K) = c(K)$, where $c(K)$ is the crossing number, that is, the minimum number of crossings among all plane diagrams representing the knot K . Besides, Moffat (1990) suggests that the minimum energy spectrum of a magnetic knot can be used to construct new knot invariants.

Topological Changes, Dissipation, and Reconnection in Fluid Patterns

As we saw above, topological changes do occur when dissipative effects become predominant over the coherency of structures. When this happens, there is a dramatic change of fluid patterns, often on small timescales compared to evolution. The change occurs through the formation and disappearance of physical reconnections in the fluid pattern. In real fluids, for example, vortex and magnetic tubes do interact and reconnect freely. From a dynamical system viewpoint, reconnections take place when the

vector field lines (streamlines, vortex lines, or magnetic lines) cross each other. If two field lines meet, the point of crossing is a true nodal point, like a bifurcation in a path. Dissipative effects allow the reconnection to proceed through such points.

In dissipative fluids, mathematical and physical properties are no longer conserved, and during the process we lose part of the original information. However, some of the invariants are rather robust and may only degrade slowly. One of them is magnetic helicity, the magnetic analog of the kinetic helicity. Its dissipation during reconnection can be modest; in particular, if the reconnection timescale is small compared to classical dissipation times, then helicity loss will be negligible. The robustness of magnetic helicity plays a central role in fusion plasma physics and in many astrophysical contexts. On the other hand, large changes in kinetic helicity are intimately related to qualitative changes in the topology of vortex flows.

Under Euler's equations, the helicity of a vortex tube of vorticity ω and velocity \mathbf{u} is defined by $H = \int \mathbf{u} \cdot \omega dV$. The integral is taken over the tube volume V occupied by ω . Now, for n knotted and linked vortex tubes, each of (constant) strength (total vorticity) $\Phi_i (1 \leq i \leq N)$, the helicity of the whole system can be expressed in terms of linking numbers Lk_{ij} as

$$H = \sum_{ij} \text{Lk}_{ij} \Phi_i \Phi_j$$

which is equal to Lk_{ij} ; this is a topological invariant whose value does not change under continuous deformation of the fluid structure. Since helicity and flux-tube strength are measurable conserved quantities, the above equation provides useful information about the topology of the flow field and flow structures. In addition, by direct measurements of helicity and application of conservation of topology, one can estimate average geometric quantities, such as the mean twist of field lines, and their contribution to the total energy.

Brief Conclusion

In this article, we have made an attempt to indicate how "classical" field theories, which have been successfully used to describe physics of fundamental structures and forces of nature, can also be used to study geometry and topology of low-dimensional manifolds. These developments not only provide new insights into old problems of topology of these manifolds but also have been responsible for profoundly interesting new mathematics (fluid mechanics, dynamical flows, and polymer biophysics

are maybe the most significant examples in the last years). In particular fluid dynamics, a topological macroscopic field theory, provides a powerful framework for modern theory of knots and links in 3-manifolds. Moreover, as we saw here, it provides a physical interpretation of the link, self-linking, and writhing number of knots and links. The present article was essentially aimed to illustrate such a relationship. Thus, the most fundamental result we reported here is the relation (formula) connecting the helicity of vector (magnetic) fields to the writhing number of knots: $H(V) = \text{Flux}(V)^2 \text{Wr}(K)$. So, writhing number for knots is the analog of helicity for vector fields. Both expressions of these invariants are variants of the (Gaussian) integral formula for the linking number of two disjoint closed space curves. Further investigations of these invariants and their mathematical properties might throw new light on the interfaces between many different areas of macroscopic and quantum physics.

See also: The Jones Polynomial; Knot Theory and Physics; Magnetohydrodynamics; Mathematical Knot Theory; Stability of Flows; Superfluids; Topological Quantum Field Theory: Overview; Vortex Dynamics; Yang–Baxter Equations.

Further Reading

- Arnol'd V and Khesin B (1998) *Topological Methods in Hydrodynamics*. Heidelberg: Springer.
- Berger MA and Field GB (1984) The topological properties of magnetic helicity. *Journal of Fluid Mechanics* 147: 133–148.
- Berry MV and Dennis MR (2001) Knotted and linked phase singularities in nonchromatic waves. *Proceedings of the Royal Society A* 457: 2251–2263.
- Boi L (2005) Topological knots' models in physics and biology. In: Boi L (ed.) *Geometries of Nature, Living Systems and Human Cognition. New Interactions of Mathematics with Natural Sciences and Humanities*, pp. 211–294. Singapore: World Scientific.
- Boyland P (2001) Fluid mechanics and mathematical structures. In: Ricca RL (ed.) *An Introduction to the Geometry and Topology of Fluid Flows*, pp. 105–134. NATO-ASI Series: Mathematics. Dordrecht: Kluwer.
- Cantarella J, De Turk D, and Gkuck H (2001) The Biot–Savart operator for application to knot theory fluid dynamics, and plasma physics. *Journal of Mathematical Physics* 42: 876–905.
- Freedman MH and Zheng-Xu He (1991) Divergence free fields: energy and asymptotic crossing number. *Annals of Mathematics* 134: 189–229.
- Fuller FB (1978) Decomposition of the linking number of a closed ribbon: a problem from molecular biology. *Proceedings of the National Academy of Sciences, USA* 75: 3557–3561.
- Ghrist RW, Holmes PhJ, and Sullivan MC (1997) *Knots and Links in Three-Dimensional Flows*. Heidelberg: Springer.
- Hornig G (2002) *Topological Methods in Fluid Dynamics*. Preprint. Ruhr-Universität-Bochum.
- Kauffman LH (1995) *Knots and Applications. Series on Knots and Everything*, vol. 6, Singapore: World Scientific.
- Lomonaco SJ (1995) The modern legacies of Thomson's atomic vortex theory in classical electrodynamics. In: Kauffman LH (ed.) *The Interface of Knots and Physics*, Proc. Symp. Appl. Math., vol. 51, pp. 145–166. American Mathematical Society.
- Moffatt HK (1969) The degree of knottedness of tangled vortex lines. *Journal of Fluid Mechanics* 35: 117–129.
- Moffatt HK (1990) The energy spectrum of knots and links. *Nature* 347: 367–369.
- Moffatt HK, Zaslavsky GM, Comte P, and Tabor M (1992) *Topological Aspects of the Dynamics of Fluids and Plasmas*, NATO ASI Series, Series E: Applied Sciences, vol. 218. Dordrecht: Kluwer Academic.
- Ricca RL (1998) New developments in topological fluid mechanics: from Kelvin's vortex knots to magnetic knots. In: Stasiak A, Katritch V, and Kauffman LH (eds.) *Ideal Knots*. Singapore: World Scientific.
- Ricca RL and Moffatt HK (1992) The helicity of a knotted vortex filament. In: Moffatt HK (ed.) *Topological Aspects of Dynamics of Fluids and Plasmas*, pp. 225–236. Dordrecht: Kluwer.
- Tait PG (1900) On Knots I, II, III. In: *Scientific Papers*. Cambridge: Cambridge University Press.
- Thomson JJ (1883) *A Treatise on the Motion of Vortex Rings*. London: Macmillan.
- Trueba JL and Rañada AF (2000) Helicity in classical electrodynamics and its topological quantization. *Apeiron* 7: 83–88.
- Woltjer L (1958) A theorem on force-free magnetic fields. *Proceedings of the National Academy of Sciences, USA* 44: 489–491.

Topological Quantum Field Theory: Overview

J M F Labastida, CSIC, Madrid, Spain

C Lozano, INTA, Torrejón de Ardoz, Spain

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Topological quantum field theory (TQFT) constitutes one of the most successful fields of mathematical physics since it originated in the 1980s. It possesses an inherent property which makes it unique: TQFT provides predictions in mathematics which open

new fields of research. A well-known example is the prediction of Seiberg–Witten invariants as building blocks of Donaldson invariants. However, there are others such as the recent proposal for the coefficients of the HOMFLY polynomial invariants for knots as quantities related to enumerative geometry. These developments have drawn the attention of mathematicians and physicists into TQFT since the 1980s, a very fruitful period in which both communities have benefited from each other.

Topology has always been present in mathematical physics, in particular when dealing with aspects of

quantum physics. Global effects play an important role in quantum-mechanical models and topology becomes an essential ingredient in their description. TQFT itself appeared in the winter of 1987 after Witten's work (Witten 1988a) on Donaldson theory (Donaldson 1990), but a series of papers during the 1980s which dealt with topological aspects of field and string theory anticipated its existence. Two of these correspond to Witten's works on supersymmetric quantum mechanics and supersymmetric sigma models (Witten 1982) that led to a generalization of Morse theory. This generalization was considered by Floer (1987) in a new context that constituted the key element in Witten's construction of TQFT. These developments were certainly influenced by Atiyah (1988). TQFT was born as a result of the interplay between physics and mathematics. This has been a constant feature all along its development.

Soon after the formulation of the TQFT addressing Donaldson theory, now known as Donaldson–Witten theory, Witten formulated a new TQFT which focuses on knot invariants such as the Jones polynomial and its generalizations (Jones 1985). Witten (1989) constructed Chern–Simons gauge theory and proved its relation to the theory of knot and link invariants. This theory possesses different features than Donaldson–Witten theory, and in fact it turns out that these two theories fall into two different general types of TQFTs as will be explained in the following section. Anyhow, despite their formal differences, both Donaldson–Witten and Chern–Simons gauge theory emerged as a novel way to express topological invariants in terms of quantum field theory quantities as well as to generalize their previous formulation. But there was much more to them than it seemed in their beginnings. Once these topological invariants were formulated in field theory language, one had a huge machinery to study them from different points of view. Theoretical physicists have developed many useful tools to study quantum field theory. The use of these tools led to new frameworks for these topological invariants.

In this overview we are going to provide the basics of TQFT and briefly describe two examples – Donaldson–Witten theory and Chern–Simons gauge theory – to explain how the general features are implemented. Some excellent reviews on the subject (Birmingham *et al.* 1991, Cordes *et al.* 1996, Labastida and Mariño 2004) are available. The organization of this work is as follows. In the following section we present a general introduction to TQFT from a functional integral point of view. Next, we touch upon the twisting of extended supersymmetry as a general constructive approach to TQFT. This is followed by a section on

Donaldson–Witten theory where we discuss the computation of its observables from a perturbative approach, showing their relation to the Donaldson invariants. Next, we introduce Chern–Simons gauge theory as a theory of knot and link invariants. The penultimate section deals with advanced developments in TQFT. Finally, we end up with some concluding remarks.

Topological Quantum Field Theory

We will start our overview by presenting the most general structure of a TQFT from a functional integral point of view which, though not rigorously defined, is the approach that has led to the most important developments. As in conventional quantum field theory, axiomatic approaches to TQFT do exist, but we will not follow that route here.

Let us consider an n -dimensional Riemannian manifold X endowed with a metric $g_{\mu\nu}$ and a quantum field theory on it. We will say that this theory is “topological” if there exist operators in the theory such that their correlation functions do not depend on the metric. If we denote these operators by \mathcal{O}_i (where i is a generic label), then

$$\frac{\delta}{\delta g_{\mu\nu}} \langle \mathcal{O}_{i_1} \cdots \mathcal{O}_{i_n} \rangle = 0 \quad [1]$$

where $\langle \cdots \rangle$ denotes a vacuum expectation value. The operators that satisfy this equation are called “topological observables.”

The simplest way to achieve metric independence is to consider a theory whose action and operators do not depend on the metric. In this situation, if no anomalous metric dependence is generated upon quantization, the correlation functions of these operators satisfy [1] and lead to topological invariants on X . Theories of this sort are collectively referred to as Schwarz-type TQFTs, and well-known examples are Chern–Simons gauge theory and BF theories. However, Schwarz-type theories are too restrictive. One would like to have a theory satisfying property [1] with a weaker condition on the action. This can be achieved with the help of a symmetry. The resulting TQFTs are called of Witten or cohomological type, the main examples being Donaldson–Witten theory and topological sigma models (Witten 1988b).

For TQFTs of Witten type, the action may depend on the metric. However, the theory has an underlying scalar symmetry δ acting on the fields ϕ_i . Since δ is a symmetry, the action of the theory satisfies $\delta S(\phi_i) = 0$. In these theories, metric independence of the correlation functions is achieved as follows. Let $T_{\mu\nu} = (\delta/\delta g^{\mu\nu})S(\phi_i)$ be the energy–momentum tensor of

the theory. It turns out that the energy–momentum tensor is δ -exact:

$$T_{\mu\nu} = -i\delta G_{\mu\nu} \quad [2]$$

$G_{\mu\nu}$ being some tensor. Indeed, if [2] is satisfied, it follows that for any set of operators \mathcal{O}_i which are δ -invariant,

$$\begin{aligned} \frac{\delta}{\delta g^{\mu\nu}} \langle \mathcal{O}_{i_1} \mathcal{O}_{i_2} \cdots \mathcal{O}_{i_n} \rangle &= \langle \mathcal{O}_{i_1} \mathcal{O}_{i_2} \cdots \mathcal{O}_{i_n} T_{\mu\nu} \rangle \\ &= -i \langle \mathcal{O}_{i_1} \mathcal{O}_{i_2} \cdots \mathcal{O}_{i_n} \delta G_{\mu\nu} \rangle \\ &= \pm i \langle \delta(\mathcal{O}_{i_1} \mathcal{O}_{i_2} \cdots \mathcal{O}_{i_n} G_{\mu\nu}) \rangle \\ &= 0 \end{aligned} \quad [3]$$

In this computation we have assumed that the symmetry δ is not anomalous and that there are no contributions coming from boundary terms since we have integrated by parts in field space. This is not always the case and in fact the situations in which one of these two properties fails lead to rich phenomena. In those cases, for example, in Donaldson–Witten theory on manifolds with $b_2^+ = 1$, the correlation functions fail to be topological invariants in a controlled manner which unveils many interesting properties.

We will now describe Witten-type theories in a general context. The general structure of Schwarz-type theories is much simpler and will be illustrated in the example presented below. In Witten-type theories the observables are the δ -invariant operators. It is simple to prove that δ -exact operators decouple from the theory. Indeed, if \mathcal{O}_a is δ -exact, $\mathcal{O}_a = \delta \hat{\mathcal{O}}_a$, then

$$\begin{aligned} \langle \mathcal{O}_a \mathcal{O}_{i_1} \mathcal{O}_{i_2} \cdots \mathcal{O}_{i_n} \rangle &= \langle \delta \hat{\mathcal{O}}_a \mathcal{O}_{i_1} \mathcal{O}_{i_2} \cdots \mathcal{O}_{i_n} \rangle \\ &= \langle \delta(\hat{\mathcal{O}}_a \mathcal{O}_{i_1} \mathcal{O}_{i_2} \cdots \mathcal{O}_{i_n}) \rangle = 0 \end{aligned} \quad [4]$$

Thus, one can restrict the set of observables to the cohomology of δ :

$$\mathcal{O} \in \frac{\text{Ker } \delta}{\text{Im } \delta} \quad [5]$$

There is no reason *a priori* why the δ -symmetry should be a scalar Grassmannian symmetry, but in all known models of Witten-type TQFTs this turns out to be the case. Thus, these theories violate the spin-statistics theorem. In all these models the algebra of the δ symmetry has the form

$$\delta^2 = Z \quad [6]$$

where Z is a symmetry transformation (typically a gauge symmetry of some sort). This property forces to consider Z -invariant observables and to work in the context of “equivariant cohomology.”

The observables of Witten-type theories fit into a general pattern that we describe now. The key ingredient is a map between the homology of X and

the equivariant cohomology of δ . Given an operator $\phi^{(0)}$ in the equivariant cohomology of δ , let us consider the following set of equations:

$$d\phi^{(n)} = \delta\phi^{(n+1)}, \quad n \geq 0 \quad [7]$$

where the operators $\phi^{(n)} (n=1, \dots, \dim X)$ are differential forms of degree n on X and d is the de Rham differential. These differential equations are called “descent equations” and their solutions $\phi^{(n)} (n \geq 0)$ “topological descendants” of $\phi^{(0)}$. We will show how to construct a solution to these equations on general grounds.

The topological descendants lead to the construction of a set of elements of the equivariant cohomology of δ . Let γ_n be an n -cycle on X , $\gamma_n \in H_n(X)$, and let us consider the following operator:

$$W_{\phi^{(0)}}^{(\gamma_n)} = \int_{\gamma_n} \phi^{(n)} \quad [8]$$

This operator is δ -invariant,

$$\delta W_{\phi^{(0)}}^{(\gamma_n)} = \int_{\gamma_n} \delta\phi^{(n)} = \int_{\gamma_n} d\phi^{(n-1)} = \int_{\partial\gamma_n} \phi^{(n-1)} = 0 \quad [9]$$

since $\partial\gamma_n = 0$. On the other hand, if γ_n were trivial in homology, that is, if $\gamma_n = \partial\gamma_{n+1}$, we would have that $W_{\phi^{(0)}}^{(\gamma_n)}$ is δ -exact:

$$W_{\phi^{(0)}}^{(\gamma_n)} = \int_{\partial\gamma_{n+1}} \phi^{(n)} = \int_{\gamma_{n+1}} d\phi^{(n)} = \delta \int_{\gamma_{n+1}} \phi^{(n+1)} \quad [10]$$

Thus, given the operator $\phi^{(0)}$, we have constructed a map between the homology of X and the equivariant cohomology of δ . There are as many maps as basic operators $\phi^{(0)}$ one finds in the theory.

To actually construct these maps, we need to find a solution of the descent equations [7]. As announced before, there is a general solution to those equations in Witten-type theories. Since in this type of theories [2] holds, there exists an operator

$$G_\mu \equiv G_{0\mu} \quad [11]$$

that satisfies

$$P_\mu = T_{0\mu} = -i\delta G_\mu \quad [12]$$

Notice that G_μ is an anticommuting operator and a 1-form in spacetime. With the aid of this operator, one constructs the following solution to the descent equations [7]:

$$\phi^{(n)} = \frac{1}{n!} \phi_{\mu_1 \mu_2 \dots \mu_n}^{(n)} dx^{\mu_1} \wedge \cdots \wedge dx^{\mu_n} \quad [13]$$

where

$$\begin{aligned} \phi_{\mu_1 \mu_2 \dots \mu_n}^{(n)}(x) &= G_{\mu_1} G_{\mu_2} \cdots G_{\mu_n} \phi^{(0)}(x), \\ n &= 1, \dots, \dim X \end{aligned} \quad [14]$$

One can easily check using [12] and the δ -invariance of $\phi^{(0)}$ that the operators [13] do satisfy the descent equations [7].

We have seen that Witten-type TQFTs are characterized by property [2]. It would be desirable to have at hand a systematic procedure to build theories satisfying that property. It has been found that extended supersymmetry provides a very helpful starting point to build those theories. Although supersymmetry guarantees from first principles only the weaker condition [12] instead of [2], all TQFTs that have been constructed from extended supersymmetry actually satisfy [2]. To build a TQFT from a theory with extended supersymmetry, one needs to go through the twisting procedure that we now describe.

Twisting of Extended Supersymmetry

All known Witten-type theories are related to an underlying extended supersymmetric quantum field theory. The topological theory is a modified version of the supersymmetric theory in which the Lorentz transformation properties (spins) of some of the fields have been modified. This modification of spin assignments is known as twisting, and it can be carried out on any theory with extended supersymmetry in any spacetime dimension. We will not consider the procedure in such a general setting but instead we will illustrate it by considering the case of $\mathcal{N}=2$ supersymmetry in four dimensions. We will begin with a general description and then we will apply it to a specific example: Donaldson–Witten theory.

Let us consider the Euclidean version of the $\mathcal{N}=2$ supersymmetry algebra with no central charges. Central charges can be included without much ado but we will not consider them for simplicity. The total symmetry group of the theory is $\mathcal{H} = \text{SU}(2)_+ \times \text{SU}(2)_- \times \text{SU}(2)_R \times \text{U}(1)_R$, $\mathcal{K} = \text{SU}(2)_+ \times \text{SU}(2)_-$ being the rotation group, and $\text{SU}(2)_R \times \text{U}(1)_R$ the internal symmetry group of the $\mathcal{N}=2$ supersymmetry algebra. The generator algebra takes the following form:

$$\begin{aligned} \{Q_{\alpha\nu}, \bar{Q}_{\dot{\beta}w}\} &= 2\epsilon_{\nu w}\sigma_{\alpha\dot{\beta}}^\mu P_\mu, & \{Q_{\alpha\nu}, Q_{\beta w}\} &= 0 \\ [P_\mu, Q_{\alpha\nu}] &= 0, & [P_\mu, \bar{Q}_{\dot{\alpha}w}] &= 0 \\ [M_{\alpha\beta}, Q_{\delta\nu}] &= \epsilon_{\delta(\alpha}Q_{\beta)\nu}, & [M_{\alpha\beta}, \bar{Q}_{\dot{\alpha}w}] &= 0 \\ [\bar{M}_{\dot{\alpha}\dot{\beta}}, Q_{\delta\nu}] &= 0, & [\bar{M}_{\dot{\alpha}\dot{\beta}}, \bar{Q}_{\dot{\delta}w}] &= \epsilon_{\dot{\delta}(\dot{\alpha}}\bar{Q}_{\dot{\beta})w} \\ [B^{\nu w}, Q_\alpha] &= \epsilon^{u(\nu}Q_\alpha^{w)}, & [B^{\nu w}, \bar{Q}_{\dot{\alpha}}] &= -\epsilon^{u(\nu}\bar{Q}_{\dot{\alpha}}^{w)} \\ [Q_{\alpha\nu}, R] &= Q_{\alpha\nu}, & [\bar{Q}_{\dot{\alpha}w}, R] &= -\bar{Q}_{\dot{\alpha}w} \end{aligned} \quad [15]$$

In these relations $\nu, w \in \{1, 2\}$ are $\text{SU}(2)_R$ indices and α and $\dot{\alpha}$ denote spinorial indices of $\text{SU}(2)_-$ and $\text{SU}(2)_+$, respectively. The supersymmetry generators

$Q_{\alpha\nu}$ and $\bar{Q}_{\dot{\alpha}w}$ transform under \mathcal{H} as $(0, 2, 2)^1$ and $(2, 0, 2)^{-1}$, respectively. $\bar{M}_{\dot{\alpha}\dot{\beta}}$ and $M_{\alpha\beta}$ are the generators of $\text{SU}(2)_+$ and $\text{SU}(2)_-$, respectively, while $B^{\nu w}$ and R generate $\text{SU}(2)_R$ and $\text{U}(1)_R$, respectively.

The twisting of a supersymmetric theory involves a modification of the couplings of the theory to a background metric on the space where the theory is defined. This modification is carried out redefining the Lorentz transformation properties of the different fields making use of the internal symmetry $\text{SU}(2)_R$. In particular, we will redefine the couplings of the fields to the $\text{SU}(2)_+$ spin connection according to the way they transform under $\text{SU}(2)_R$. This is easily done by identifying the $\text{SU}(2)_R$ indices ν with the $\text{SU}(2)_+$ indices $\dot{\alpha}$. The procedure involves a redefinition of the rotation group into $\mathcal{K}' = \text{SU}'(2)_+ \otimes \text{SU}(2)_-$, where $\text{SU}'(2)_+$ is generated by

$$M'_{\dot{\alpha}\dot{\beta}} = M_{\dot{\alpha}\dot{\beta}} - B_{\dot{\alpha}\dot{\beta}} \quad [16]$$

The supersymmetry generators $Q_{\alpha\nu}$ and $\bar{Q}_{\dot{\alpha}w}$ get transformed in the following way:

$$\begin{aligned} \bar{Q}_{\dot{\alpha}w} &\rightarrow \bar{Q}_{\dot{\alpha}\dot{\beta}} \\ Q_{\alpha\nu} &\rightarrow Q_{\alpha\dot{\beta}} \end{aligned} \quad [17]$$

which allows us to define the “topological supercharge”:

$$\bar{Q} \equiv \epsilon^{\dot{\alpha}\dot{\beta}} \bar{Q}_{\dot{\alpha}\dot{\beta}} \quad [18]$$

It is simple to prove using [15] and [16] that this quantity is a scalar under the new rotation group \mathcal{K}' : $[M_{\alpha\beta}, \bar{Q}] = 0$ and $[M'_{\dot{\alpha}\dot{\beta}}, \bar{Q}] = 0$. In addition, from [15], it follows that \bar{Q} is nilpotent (in the absence of central charges):

$$\bar{Q}^2 = 0 \quad [19]$$

The scalar generator \bar{Q} leads to the topological symmetry δ of the previous section. Actually, the twisting procedure provides also the operator G_μ in [12]. Defining

$$G_\mu = \frac{i}{4}(\bar{\sigma}_\mu)^{\dot{\alpha}\gamma} Q_{\gamma\dot{\alpha}} \quad [20]$$

one easily finds, after using [15] and [18],

$$\{\bar{Q}, G_\mu\} = \partial_\mu \quad [21]$$

which is indeed equivalent to [12]. On general grounds we cannot prove that twisted supersymmetric theories lead to theories which satisfy [12]. However relation [12], which is weaker, is guaranteed. It turns out that in all the models originated from extended supersymmetry which have been studied, [2] is satisfied and thus the resulting theories are TQFTs of Witten type.

Donaldson–Witten Theory

One of the greatest successes of TQFT has been the discovery of Seiberg–Witten invariants as building blocks of Donaldson invariants. This was achieved in two main steps. First, Donaldson theory was reformulated in field-theoretical terms, using perturbative methods. Second, the resulting TQFT was solved using nonperturbative methods. In this section we are going to describe in some detail the first step. The second one will be briefly addressed later and is the main object of a separate article in the encyclopedia (*see* Seiberg–Witten Theory).

Let us consider $\mathcal{N}=2$ supersymmetric Yang–Mills theory in four dimensions. The field content of the theory is the following: a gauge field A_μ , two spinors $\lambda_{v\alpha}$, and a complex scalar ϕ , all of them in the adjoint representation of a gauge group \mathcal{G} . In addition, the theory possesses the auxiliary fields D_{vw} in the 3 of the internal $SU(2)_R$. The theory has the following action:

$$\begin{aligned} \int d^4x \operatorname{tr} \left(\nabla_\mu \phi^\dagger \nabla^\mu \phi - i \lambda_v \sigma^\mu \nabla_\mu \bar{\lambda}^v - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} \right. \\ \left. + \frac{1}{4} D_{vw} D^{vw} - \frac{1}{2} [\phi, \phi^\dagger]^2 - \frac{i}{\sqrt{2}} \epsilon^{vw} \lambda_v^\alpha [\phi^\dagger, \lambda_{w\alpha}] \right. \\ \left. - \frac{i}{\sqrt{2}} \epsilon_{vw} \bar{\lambda}^v_{\dot{\alpha}} [\bar{\lambda}^{w\dot{\alpha}}, \phi] \right) \end{aligned} \quad [22]$$

This action is invariant under the following $\mathcal{N}=2$ supersymmetric transformations:

$$\begin{aligned} \delta \phi &= \sqrt{2} \epsilon^{vw} \xi_v \lambda_w \\ \delta A_\mu &= i \xi_v \sigma_\mu \bar{\lambda}^v - i \lambda_v \sigma_\mu \bar{\xi}^v \\ \delta \lambda_{v\alpha} &= D_v{}^w \xi_{w\alpha} - i \xi_{v\alpha} [\phi, \phi^\dagger] - i \sigma^{\mu\nu}{}_\alpha{}^\beta \xi_{v\beta} F_{\mu\nu} \\ &\quad + i \sqrt{2} \epsilon_{vw} \sigma_{\alpha\dot{\alpha}}^\mu \bar{\xi}^{w\dot{\alpha}} \nabla_\mu \phi \\ \delta D^{vw} &= 2i \bar{\xi}^{(v} \bar{\sigma}^\mu \nabla_\mu \lambda^{w)} + 2i \nabla_\mu \bar{\lambda}^{(v} \bar{\sigma}^\mu \xi^{w)} \\ &\quad + 2i \sqrt{2} \xi^{(v} [\lambda^{w)}, \phi^\dagger] + 2i \sqrt{2} \bar{\xi}^{(v} [\bar{\lambda}^{w)}, \phi] \end{aligned} \quad [23]$$

ξ_v being spinorial $\mathcal{N}=2$ supersymmetric parameters.

We can now twist the above theory following the procedure explained in the previous section. Upon twisting, the fields of the theory change their spin content as follows:

$$\begin{aligned} A_\mu(2, 2, 0)^0 &\rightarrow A_\mu(2, 2)^0 \\ \lambda_{\alpha v}(2, 0, 2)^1 &\rightarrow \psi_{\alpha\dot{\beta}}(2, 2)^1 \\ \bar{\lambda}_{\dot{\alpha} v}(0, 2, 2)^{-1} &\rightarrow \eta(0, 0)^{-1}, \chi_{\dot{\alpha}\dot{\beta}}(3, 0)^{-1} \\ \phi(0, 0, 0)^2 &\rightarrow \phi(0, 0)^2 \\ \phi^\dagger(0, 0, 0)^{-2} &\rightarrow \phi^\dagger(0, 0)^{-2} \\ D_{vw}(0, 0, 3)^0 &\rightarrow D_{\dot{\alpha}\dot{\beta}}(0, 0)^0 \end{aligned} \quad [24]$$

In this table the representations of the respective rotation groups carried by the fields have been indicated. The superindices refer to the $U(1)_R$ charge which is also called “ghost number” in the context of TQFT. The fields η and χ are given by the antisymmetric and symmetric pieces of $\bar{\lambda}_{\dot{\alpha}\dot{\beta}}: \chi_{\dot{\alpha}\dot{\beta}} = \bar{\lambda}_{(\dot{\alpha}\dot{\beta})}$ and $\eta = (1/2) \epsilon^{\dot{\alpha}\dot{\beta}} \bar{\lambda}_{\dot{\alpha}\dot{\beta}}$.

Notice that the twisted fields in [24] are differential forms on X ; therefore, the twisted theory makes sense globally on any arbitrary Riemannian 4-manifold. This is not the case with the original $\mathcal{N}=2$ supersymmetric Yang–Mills, which contains fermionic fields. Making global sense of those on arbitrary Riemannian 4-manifolds requires the manifold to be Spin.

The dynamics of the twisted theory is governed by an action which can be obtained by twisting the action [22]. On an arbitrary Riemannian 4-manifold endowed with a metric $g_{\mu\nu}$, the twisted action becomes

$$\begin{aligned} S = \int d^4x \sqrt{g} \operatorname{tr} \left(\nabla_\mu \phi \nabla^\mu \phi^\dagger - i \psi_{\alpha\dot{\beta}} \sigma^{\mu\dot{\alpha}\alpha} \nabla_\mu \chi_{\dot{\alpha}\dot{\beta}} \right. \\ \left. - i \psi_{\alpha\dot{\alpha}} \nabla^{\dot{\alpha}\alpha} \eta - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \frac{1}{4} D_{\dot{\alpha}\dot{\beta}} D^{\dot{\alpha}\dot{\beta}} \right. \\ \left. - \frac{1}{2} [\phi, \phi^\dagger]^2 - \frac{i}{\sqrt{2}} \chi^{\dot{\alpha}\dot{\beta}} [\phi, \chi_{\dot{\alpha}\dot{\beta}}] \right. \\ \left. + i \sqrt{2} \eta [\phi, \eta] - \frac{i}{\sqrt{2}} \psi_{\alpha\dot{\alpha}} [\psi^{\alpha\dot{\alpha}}, \phi^\dagger] \right) \end{aligned} \quad [25]$$

where $\sqrt{g} = (\det(g_{\mu\nu}))^{1/2}$.

To obtain the transformations of the fields under the topological symmetry, we need to compute the $\bar{\mathcal{Q}}$ -transformations. These are easily obtained using [18] and [23]. They turn out to be

$$\begin{aligned} [\bar{\mathcal{Q}}, \phi] &= 0 \\ [\bar{\mathcal{Q}}, A_\mu] &= \psi_\mu \\ \{\bar{\mathcal{Q}}, \eta\} &= [\phi, \phi^\dagger] \\ \{\bar{\mathcal{Q}}, \psi_\mu\} &= 2\sqrt{2} \nabla_\mu \phi \\ [\bar{\mathcal{Q}}, \phi^\dagger] &= 2\sqrt{2} i \eta \\ \{\bar{\mathcal{Q}}, \chi_{\dot{\alpha}\dot{\beta}}\} &= i(F_{\dot{\alpha}\dot{\beta}}^+ - D_{\dot{\alpha}\dot{\beta}}) \\ [\bar{\mathcal{Q}}, D] &= (2\nabla\psi)^+ + 2\sqrt{2} [\phi, \chi] \end{aligned} \quad [26]$$

where $\psi_\mu = \sigma_{\mu\alpha\dot{\beta}} \psi^{\alpha\dot{\beta}}$ and $F_{\dot{\alpha}\dot{\beta}}^+ = \sigma_{\dot{\alpha}\dot{\beta}}^{\mu\nu} F_{\mu\nu}$ is the self-dual part of $F_{\mu\nu}$. Using these transformations, one easily finds that $\bar{\mathcal{Q}}^2$ is a gauge transformation. This is not unexpected since the $\mathcal{N}=2$ supersymmetric transformations [23] are in the Wess–Zumino gauge and they close only up to gauge transformations. This property implies that one must consider the equivariant cohomology of $\bar{\mathcal{Q}}$ defined on the set of gauge-invariant operators.

The action [25] is $\overline{\mathcal{Q}}$ -exact up to a topological term:

$$S = \{\overline{\mathcal{Q}}, V\} - \frac{1}{2} \int F \wedge F \quad [27]$$

where

$$V = \int d^4x \sqrt{g} \operatorname{tr} \left(\frac{i}{4} \chi_{\dot{\alpha}\dot{\beta}} (F^{\dot{\alpha}\dot{\beta}} + D^{\dot{\alpha}\dot{\beta}}) - \frac{1}{2} \eta[\phi, \phi^\dagger] + \frac{1}{2\sqrt{2}} \psi_{\alpha\dot{\alpha}} \nabla^{\dot{\alpha}\alpha} \phi^\dagger \right) \quad [28]$$

Actually, it turns out that in all the theories obtained after twisting extended supersymmetry, the resulting actions are $\overline{\mathcal{Q}}$ -exact up to topological terms. In the case of $\mathcal{N}=2$ theories, topological (theta) terms $\int F \wedge F$ are generically not observable (due to a chiral anomaly), so it is customary to pick

$$S_{\text{DW}} = \{\overline{\mathcal{Q}}, V\} \quad [29]$$

as the action of the theory, which immediately implies [2] and therefore the topological character of the theory. Notice, however, that [29] is stronger than [2].

As we described in the previous section, the observables of the theory can be constructed using the operator G_μ in [20]. Its action on the twisted fields is easily obtained using [23]:

$$\begin{aligned} [G_\mu, \phi] &= \frac{1}{2\sqrt{2}} \psi_\mu \\ [G_\nu, A_\mu] &= \frac{i}{2} g_{\mu\nu} \eta - i \chi_{\mu\nu} \\ [G, \eta] &= -\frac{i\sqrt{2}}{4} \nabla \bar{\phi} \\ \{G_\mu, \psi_\nu\} &= -(F_{\mu\nu}^- + D_{\mu\nu}^+) \\ [G, \bar{\phi}] &= 0 \\ [G, F^+] &= i \nabla \chi + \frac{3i}{2} * \nabla \eta \\ \{G, \chi\} &= -\frac{3i\sqrt{2}}{8} * \nabla \bar{\phi} \\ [G, D] &= -\frac{3i}{4} * \nabla \eta + \frac{3i}{2} \nabla \chi \end{aligned} \quad [30]$$

We now need to fix the basic operator $\phi^{(0)}$ in [14]. The starting point must be a set of gauge-invariant, $\overline{\mathcal{Q}}$ -closed operators which are not $\overline{\mathcal{Q}}$ -trivial. Since $[\overline{\mathcal{Q}}, \phi] = 0$, these operators are the gauge-invariant polynomials in the field ϕ . For a simple gauge group of rank r the algebra of these polynomials is generated by r elements, and we shall denote this basis by $\mathcal{O}_n, n=1, \dots, r$. A simple choice for $\text{SU}(N)$ consists of the following Casimirs:

$$\mathcal{O}_n = \operatorname{tr}(\phi^{n+1}), \quad n = 1, \dots, N \quad [31]$$

Using G_μ we can now construct the map between the homology of X and the equivariant cohomology of $\overline{\mathcal{Q}}$. Let us consider the simple case $\text{SU}(2)$. There exists only one independent Casimir and, correspondingly, only one basic operator:

$$\mathcal{O} = \operatorname{tr}(\phi^2) \quad [32]$$

for which one finds the following set of descendants:

$$\begin{aligned} \mathcal{O}^{(1)} &= \operatorname{tr} \left(\frac{1}{\sqrt{2}} \phi \psi_\mu \right) dx^\mu \\ \mathcal{O}^{(2)} &= -\frac{1}{2} \operatorname{tr} \left(\frac{1}{\sqrt{2}} \phi (F_{\mu\nu}^- + D_{\mu\nu}) - \frac{1}{4} \psi_\mu \psi_\nu \right) dx^\mu \wedge dx^\nu \\ &\vdots \end{aligned} \quad [33]$$

The map from the homology of X to the equivariant cohomology of $\overline{\mathcal{Q}}$ can now be constructed very easily. Let γ_i be an element of the homology group $H_i(X)$. We associate to it the following observable:

$$\gamma_i \rightarrow I_i(\gamma_i) = \int_{\gamma_i} \mathcal{O}^{(i)} \quad [34]$$

where $\mathcal{O}^{(i)}$ is given in [33]. The construction assures that $I_i(\gamma_i)$ is invariant under $\overline{\mathcal{Q}}$ and gauge transformations. Furthermore, it is also assured that $I_i(\gamma_i)$ is not $\overline{\mathcal{Q}}$ -exact.

Let us consider the computation of correlation functions. The discussion will be presented for a generic gauge group. We will consider the topological theory defined by the Donaldson–Witten action

$$S_{\text{DW}} = \{\overline{\mathcal{Q}}, V\} \quad [35]$$

where V is defined in [28]. The property [35] has a very important consequence. The action S_{DW} shows up in the correlation functions as $\exp(-S_{\text{DW}}/e^2)$, where e is a free parameter which corresponds to the coupling constant of the $\mathcal{N}=2$ theory. Since the term involving the coupling constant is $\overline{\mathcal{Q}}$ -exact, the correlation functions of $\overline{\mathcal{Q}}$ -invariant operators are independent of e . Let us explain this in some detail. The (unnormalized) correlation functions of the theory are defined by

$$\langle \phi_1 \cdots \phi_n \rangle = \int \mathcal{D}\phi \phi_1 \cdots \phi_n e^{-(1/e^2) S_{\text{DW}}} \quad [36]$$

where ϕ_1, \dots, ϕ_n are invariant under $\overline{\mathcal{Q}}$ transformations. Using the fact that S_{DW} is $\overline{\mathcal{Q}}$ -exact, one obtains

$$\begin{aligned} \frac{\partial}{\partial e} \langle \phi_1 \cdots \phi_n \rangle &= \frac{2}{e^3} \langle \phi_1 \cdots \phi_n S_{\text{DW}} \rangle \\ &= \frac{2}{e^3} \langle \{\overline{\mathcal{Q}}, \phi_1 \cdots \phi_n V\} \rangle = 0 \end{aligned} \quad [37]$$

where we have used the fact that \bar{Q} is a symmetry of the theory, and therefore as in [3] the last functional integral gives zero. This result implies that one can compute these correlation functions in different limits of e . In the weak-coupling limit (semiclassical or saddle point approximation), one establishes the connection with Donaldson theory. In the strong-coupling limit, Seiberg–Witten invariants appear and one finds the connection between these two types of invariants. We will briefly explore the weak-coupling limit $e \rightarrow 0$. The functional integral [36] can be evaluated exactly in two steps: first one analyzes the zero modes or classical configurations that minimize the action, then one expands around them considering only quadratic fluctuations. The integration over these quadratic fluctuations involves ratios of determinants of kinetic operators that because of the \bar{Q} -symmetry of the theory (which in fact is a Bose–Fermi symmetry) are ± 1 . One is then left with an integral over the bosonic zero modes which leads to a finite-dimensional integral over the space of bosonic collective coordinates, and a finite Grassmannian integral over the zero modes of the fermionic fields. A careful analysis of the zero modes, first carried out by Witten, reveals that the infinite-dimensional functional integral is replaced by a finite-dimensional integral over the moduli space of anti-self-dual (ASD) connections \mathcal{M}_{ASD} , that is, the space of connections satisfying $F_{\mu\nu}^+ = 0$.

Therefore, the correlation functions [36] have the form

$$\langle \phi_1 \cdots \phi_n \rangle = \int_{\mathcal{M}_{\text{ASD}}} \hat{\phi}_1 \wedge \cdots \wedge \hat{\phi}_n \quad [38]$$

where the fields in $\phi_1 \cdots \phi_n$ are mapped to differential forms $\hat{\phi}_1 \cdots \hat{\phi}_n$ on \mathcal{M}_{ASD} – the degree of each form being given by the ghost number of its partner. Notice that the integral on the right-hand side vanishes unless the form has top degree. From the field-theoretical point of view, this is the requirement that the overall ghost number of the correlation function must be equal to $\dim \mathcal{M}_{\text{ASD}}$.

The quantities on the right-hand side of [38] are – for gauge group $\text{SU}(2)$ – precisely the Donaldson invariants. Thus, Witten’s work provided a new point of view on these invariants by reformulating them in a quantum field theory language. This is a very important contribution since quantum field theory is a very rich framework and a wide variety of methods can be used to analyze the correlation functions. This opened an entirely new strategy to investigate the Donaldson invariants. The emergence of Seiberg–Witten invariants is perhaps the greatest achievement of the implementation of this strategy.

We finish this section by pointing out that many features of the evaluation of the functional integral of the Donaldson–Witten theory developed here are common to most topological field theories of the Witten type. These features can be studied in the context of the Mathai–Quillen formalism which is the object of a separate article in the encyclopedia (see Mathai–Quillen Formalism).

Chern–Simons Gauge Theory for Knots and Links

Chern–Simons gauge theory is the most important example of Schwarz-type TQFTs. Let us begin by introducing its basic elements. Chern–Simons gauge theory is a quantum field theory whose action is based on the Chern–Simons form associated to a nonabelian gauge group. The theory is defined by the following data: a smooth 3-manifold M which will be taken to be compact, a gauge group G which will be taken semisimple and compact, and an integer parameter k . The action of the theory is

$$S_{\text{CS}}(A) = \frac{k}{4\pi} \int_M \text{tr} \left(A \wedge dA + \frac{2}{3} A \wedge A \wedge A \right) \quad [39]$$

where A is a gauge connection and the trace is taken in the fundamental representation. The exponential of i times this action is invariant under gauge transformations,

$$A \rightarrow A + g^{-1} dg \quad [40]$$

where g is a map $g: M \rightarrow G$.

Notice that the action [39] is independent of the metric on the 3-manifold M . In this theory, appropriate observables lead to correlation functions which correspond to topological invariants. Candidates to be observables of this type must be metric independent and gauge invariant. Wilson loops satisfy these properties. They correspond to the holonomy of the gauge connection A along a loop. Given a representation R of the gauge group G and a 1-cycle γ on M , it is defined as

$$W_{\gamma}^R(A) = \text{tr}_R(\text{Hol}_{\gamma}(A)) = \text{tr}_R P \exp \int_{\gamma} A \quad [41]$$

Products of these operators are the natural candidates to obtain topological invariants after computing their correlation functions. These correlation functions are formally written as

$$\begin{aligned} & \langle W_{\gamma_1}^{R_1} W_{\gamma_2}^{R_2} \cdots W_{\gamma_n}^{R_n} \rangle \\ &= \int [DA] W_{\gamma_1}^{R_1}(A) W_{\gamma_2}^{R_2}(A) \cdots W_{\gamma_n}^{R_n}(A) e^{iS_{\text{CS}}(A)} \end{aligned} \quad [42]$$

where $\gamma_1, \gamma_2, \dots, \gamma_n$ are 1-cycles on M and R_1, R_2, \dots, R_n are representations of G . In [42], the quantity $[DA]$ denotes the functional integral measure and it is assumed that an integration over connections modulo gauge transformations is carried out. As usual in quantum field theory, this integration is not well defined. Field theorists have developed methods to assign a meaning to the right-hand side of [42]. These methods mainly fall into two categories – perturbative and nonperturbative – and their degree of success mostly depends on the quantum field theory under consideration. For gauge theories, it is also possible to take an alternative approach, the large- N expansion, which in general provides further insights into the theory. In Chern–Simons gauge theory all these three methods have proved of great value.

Witten (1989) showed, using nonperturbative methods, that when one considers nonintersecting cycles $\gamma_1, \gamma_2, \dots, \gamma_n$ without self-intersections, the correlation functions [42] lead to the polynomial invariants of knot theory discovered a few years earlier starting with the work of Jones (1985).

Knot theory studies embeddings $\gamma: S^1 \rightarrow M$. Any two of such embeddings are considered equivalent if the image of one of them can be deformed into the image of the other by a homeomorphism on M . The main goal of knot theory is to classify the resulting equivalence classes. Each of these classes is a knot. Most of the work on knot theory has been carried out for the simple case $M = S^3$. Chern–Simons gauge theory, however, being a formulation intrinsically three dimensional, provides a framework to study the case of more general 3-manifolds M .

A powerful approach to classify knots is based on the construction of knot invariants. These are quantities which can be computed for a representative of a class and are invariant within the class, that is, under continuous deformations of the chosen representative. At present, it is not known if there exist enough knot invariants to classify knots. Vassiliev invariants (Vassiliev 1990) are the most promising candidates, but it is already known that if they do provide such a classification, infinitely many of them are needed.

The problem of the classification of knots in S^3 can be reformulated in a two-dimensional framework using regular knot projections. Given a representative of a knot in S^3 , deform it continuously in such a way that the projection on a plane has simple crossings. Draw the projection on the plane, and at each crossing use the convention that the line that goes under the crossing is erased in a neighborhood of the crossing. The resulting diagram is a set of segments on the plane, containing the relevant

information at the crossings. The problem of classifying knots is equivalent to the problem of classifying knot projections modulo a series of relations among them. These relations are known as Reidemeister moves. Invariance of a quantity under the three Reidemeister moves is called invariance under ambient isotopy. If a quantity is invariant under all but the first move, it is said to possess invariance under regular isotopy.

The formalism described for knots generalizes to the case of links. For a link of n components, one considers n embeddings, $\gamma_i: S^1 \rightarrow M$ ($i = 1, \dots, n$), with no intersections among them. Again, the main problem that link theory faces is the problem of their classification modulo homeomorphisms on M . In this case one can also define regular projections and reformulate the problem in terms of their classification modulo the Reidemeister moves.

The study of knot and link invariants experimented important progress in the 1980s. Jones (1985) discovered a new invariant which carries his name. The Jones polynomial can be defined very simply in terms of skein relations. These are a set of rules that can be applied to the diagram of a regular knot projection to construct the polynomial invariant. They establish a relation between the invariants associated to three links which only differ in a region as shown in Figure 1 where arrows have been introduced to take into account that the Jones polynomial is defined for oriented links.

If one denotes by $V_L(t)$ the Jones polynomial corresponding to a link L , t being the argument of the polynomial, it must satisfy the skein relation:

$$\frac{1}{t} V_{L_+} - t V_{L_-} = \left(\sqrt{t} - \frac{1}{\sqrt{t}} \right) V_{L_0} \quad [43]$$

where L_+, L_- , and L_0 are the links shown in Figure 1. This relation plus a choice of normalization for the unknot (U) are enough to compute the Jones polynomial for any link. The standard choice for the unknot is

$$V_U = 1 \quad [44]$$

though it is not the most natural one from the point of view of Chern–Simons gauge theory. After Jones

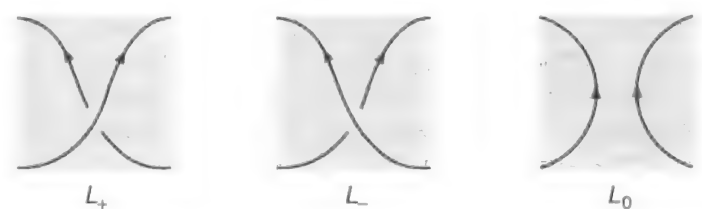


Figure 1 Skein relations.

work in 1984, many other polynomial invariants were discovered, as the HOMFLY and the Kauffman polynomial invariants.

The pioneering work of Witten in 1988 showed that the correlation functions of products of Wilson loops [42] correspond to the Jones polynomial when one considers $SU(2)$ as gauge group and all the Wilson loops entering in the correlation function are taken in the fundamental representation F . For example, if one considers a knot K , Witten showed that

$$V_K(t) = \langle W_K^F \rangle \quad [45]$$

provided that one performs the identification

$$t = \exp\left(\frac{2\pi i}{k+h}\right) \quad [46]$$

where $h=2$ is the dual Coxeter number of the gauge group $SU(2)$. Witten also showed that if instead of $SU(2)$ one considers $SU(N)$ and the Wilson loop carries the fundamental representation, the resulting invariant is the HOMFLY polynomial. The second variable of this polynomial originates in this context from the N dependence. However, these cases are just a sample of the general framework intrinsic to Chern–Simons gauge theory. Taking other groups and other representations, one possesses an enormous set of knot and link invariants. These invariants can also be obtained in the context of quantum groups.

Many nonperturbative studies of Chern–Simons gauge theory have been carried out. The quantization of the theory has been studied from the point of view of the operator formalism as well as other more geometrical methods. Also, its connection to two-dimensional conformal field theory has been further elucidated, and a powerful method for the general computation of knot and link invariants has been developed by Kaul and collaborators.

Chern–Simons theory is also amenable to perturbative analysis, which has provided important representations of the Vassiliev invariants. These invariants, proposed by Vassiliev in 1990, turned out to be the coefficients of the perturbative series expansion of the correlators of Chern–Simons gauge theory. Perturbative studies can be carried out in different gauges, originating a variety of new representations of Vassiliev invariants. Among the most relevant results related to these topics are the integral expressions for Vassiliev invariants by Kontsevich and by Bott and Taubes, as well as the recent combinatorial ones. These developments are not described here but the interested reader is referred to the recent review (Labastida 1999).

Advanced Developments

Topological sigma models are another important type of (Witten-type) TQFTs. These theories are obtained after twisting 2D $\mathcal{N}=2$ supersymmetric sigma models. The twisting can be done in two different ways leading to two types of models, A and B. Their existence is related to mirror symmetry. Only type-A models will be described in what follows. These models can be defined on an arbitrary almost-complex manifold, though typically they are considered on Kähler manifolds. The theory involves maps from two-dimensional Riemann surfaces Σ to target spaces X , together with fermionic degrees of freedom on Σ which are mapped to tangent vectors on X . The functional integral of the resulting theory is localized on holomorphic maps, defining the corresponding moduli space. The corresponding \bar{Q} -cohomology provides the set of physical observables, which can be mapped to cohomology classes on the moduli space and integrated to produce topological invariants.

Topological sigma models keep fixed the complex structure of the Riemann surface Σ . Motivated by string theory, one also considers the situation in which one integrates over complex structures. In this case, one ends up working with holomorphic maps in the entire moduli space of curves. The resulting theories are called topological strings.

We will review now a particular example of topological string theory which, besides being very interesting from the point of view of physics and mathematics, will be very useful in establishing a relation with Chern–Simons gauge theory. Let us consider topological strings with target manifold X a Calabi–Yau 3-fold. In this case, the virtual dimension of the moduli space of holomorphic maps turns out to be zero. Two situations can occur: either the space is given by a number of points (the real dimension is zero) or the moduli space is finite dimensional and possesses a bundle of the same dimension as the tangent bundle. In the first case, topological strings count the number of points weighted by the exponential of the area of the holomorphic map (the pullback of the Kähler form integrated over the surface) times x^{2g-2} , where x is the string-coupling constant and g is the genus of Σ . In the second case, one computes the top Chern class of the appropriate bundles (properly defined), again weighted by the same factor. In both cases one can classify the contributions according to the cohomology class β on X in which the image of the holomorphic map is contained. The sum of the numbers obtained for each β and fixed g are known as Gromov–Witten

invariants, N_g^β . The topological string contribution takes the form

$$\sum_{g \geq 0} x^{2g-2} \left(\sum_{\beta \in H_2(X, \mathbb{Z})} N_g^\beta e^{\int_\beta \omega} \right) \quad [47]$$

where ω is the Kähler class of the Calabi–Yau manifold. In general, the quantities N_g^β are rational numbers.

The precedent discussion has shown how Gromov–Witten invariants can be interpreted in terms of string theory. One could think that this is just a fancy observation and that no further insight on these invariants can be gained from this formulation. The situation turns out to be quite the opposite. Once a string formulation has been obtained, the whole machinery of string theory is at our disposal. One should look to new ways to compute the quantity [47], where Gromov–Witten invariants are packed. The hope is that, if this is possible, the new emerging picture will provide new insights on these invariants. This is indeed what occurred recently. It turns out that the quantity [47] can be obtained from an alternative point of view in which the embedded Riemann surfaces are regarded as D-branes. The outcome of this approach is that the Gromov–Witten invariants can be written in terms of other invariants which are integers and that possess a geometrical interpretation. To be more specific, the quantity [47] takes the form

$$\sum_{\substack{g \geq 0 \\ \beta \in H_2(X, \mathbb{Z})}} \sum_{d > 0} n_g^\beta \frac{1}{d} \left(2 \sin \left(\frac{dx}{2} \right) \right)^{2g-2} e^{d \int_\beta \omega} \quad [48]$$

where n_g^β are the new “integer” invariants. This prediction has been verified in all the cases in which it has been tested. A similar structure will be found in the next section in the context of knot theory in the large- N limit.

Let us now consider also Donaldson–Witten theory from a new perspective. To be more specific, let us consider the case in which the gauge group is $SU(2)$, and the 4-manifold X is simply connected and has $b_2^+ > 1$ (the case $b_2^+ = 1$ is anomalous). In this situation there are $1 + b_2$ physical observables [34], $\mathcal{O} = I_1$ and $I(\Sigma_a) = I_2(\Sigma_a)$ ($a = 1, \dots, b_2$), where Σ_a is a basis of $H_2(X)$. These can be packed in a generating functional:

$$\left\langle \exp \left(\sum_a \alpha_a I(\Sigma_a) + \lambda \mathcal{O} \right) \right\rangle \quad [49]$$

where λ and α_a ($a = 1, \dots, b_2$) are parameters. In computing this quantity one can argue that the contribution is localized on the moduli space of instantons configurations and one ends up, after taking into account the selection rule dictated by the dimensionality of the moduli space, with integrations

over the moduli space of the selected forms. The resulting quantities are Donaldson invariants.

As in the case of topological sigma models one could be tempted to argue that the observation leading to a field-theoretical interpretation of Donaldson invariants does not provide any new insight. Quite on the contrary, once a field theory formulation is available, one has at his disposal a huge machinery which could lead, on the one hand, to further generalizations of the theory and, on the other hand, to new ways to compute quantities such as [49], obtaining new insights on these invariants. This is indeed what happened in the 1990s, leading to an important breakthrough in 1994 when Seiberg and Witten calculated [49] in a different way and pointed out the relation of Donaldson invariants to new integer invariants that nowadays bear their names.

The localization argument that led to the interpretation of [49] as Donaldson invariants is valid because the theory under consideration is exact in the weak-coupling limit. Actually, the topological theory under consideration is independent of the coupling constant and thus calculations in the strong-coupling limit are also exact. These types of calculations were out of reach before 1994. The situation changed dramatically after the work of Seiberg and Witten in which $\mathcal{N} = 2$ super Yang–Mills theory was solved in the strong-coupling limit. Its application to the corresponding twisted version was immediate and it turned out that Donaldson invariants can be written in terms of new integer invariants now known as Seiberg–Witten invariants (Witten 1994). The development has a strong resemblance with the one described above for topological strings: certain noninteger invariants can be expressed in terms of new integer invariants.

The Seiberg–Witten invariants are actually simpler to compute than Donaldson invariants. They correspond to partition functions of topological Yang–Mills theories where the gauge group is abelian. These contributions can be grouped into classes labeled by $x = -2c_1(L)$, where $c_1(L)$ is the first Chern class of the corresponding line bundle. The sum of contributions, each being ± 1 , for a given class x is the integer Seiberg–Witten invariant n_x . The strong-coupling analysis of topological Yang–Mills theory leads to the following expression for [49]:

$$2^{1+(1/4)(7\chi+11\sigma)} \left(e^{((\nu^2/2)+2\lambda)} \sum_x n_x e^{\nu \cdot x} + i^{\chi+\sigma/4} e^{-(\nu^2/2)-2\lambda} \sum_x n_x e^{-i\nu \cdot x} \right) \quad [50]$$

where $\nu = \sum_a \alpha_a \Sigma_a$, and χ and σ are the Euler number and the signature of the manifold X . This result matches the known structure of [49] (structure theorem of Kronheimer and Mrowka) and provides

a meaning to its unknown quantities in terms of the new Seiberg–Witten invariants. Equation [50] is a rather remarkable prediction that has been tested in many cases, and for which a general proof has been recently proposed. For a review of the subject, see Labastida and Lozano (1998).

The situation for manifolds with $b_2^+ = 1$ involves a metric dependence and has been worked out in detail (Moore and Witten 1998). The formulation of Donaldson invariants in field-theoretical terms has also provided a generalization of these invariants. This generalization has been carried out in several directions: (1) the consideration of higher-rank groups, (2) the coupling to matter fields after twisting $\mathcal{N} = 2$ hypermultiplets, and (3) the twist of theories involving $\mathcal{N} = 4$ supersymmetry.

We will now look at Chern–Simons gauge theory from the perspective that emerges from its treatment in the context of the large- N expansion. We will restrict the discussion to the case of knots on S^3 with gauge group $SU(N)$. Gauge theories with gauge group $SU(N)$ admit, besides the perturbative expansion, a large- N expansion. In this expansion correlators are expanded in powers of $1/N$ while keeping the 't Hooft coupling $t = Nx$ fixed, x being the coupling constant of the gauge theory. For example, for the free energy of the theory, one has the general form

$$F = \sum_{\substack{g \geq 0 \\ h \geq 1}} C_{g,h} N^{2-2g} t^{2g-2+h} \quad [51]$$

In the case of Chern–Simons gauge theory, the coupling constant is $x = 2\pi i/(k + N)$ after taking into account the shift in k . The large- N expansion [51] resembles a string-theory expansion and indeed the quantities $C_{g,h}$ can be identified with the partition function of a topological open string with g handles and h boundaries, with N D-branes on S^3 in an ambient six-dimensional target space T^*S^3 . This was pointed out by Witten in 1992. The result makes a connection between a topological three-dimensional field theory and the topological strings described in the previous section.

In 1998 an important breakthrough took place which provided a new approach to compute quantities such as [51]. Using arguments inspired by the AdS/CFT correspondence, Gopakumar and Vafa (1999) provided a closed-string-theory interpretation of the partition function [51]. They conjectured that the free energy F can be expressed as

$$F = \sum_{g \geq 0} N^{2-2g} F_g(t) \quad [52]$$

where $F_g(t)$ corresponds to the partition function of a topological closed-string theory on the noncompact Calabi–Yau manifold X called the resolved conifold,

$\mathcal{O}(-1) \oplus \mathcal{O}(-1) \rightarrow P^1$, t being the flux of the B -field through P^1 . The quantities $F_g(t)$ have been computed using both physical and mathematical arguments, thus proving the conjecture.

Once a new picture for the partition function of Chern–Simons gauge theory is available, one should ask about the form that the expectation values of Wilson loops could take in the new context. The question was faced by Ooguri and Vafa and they provided the answer, later refined by Labastida, Mariño, and Vafa. The outcome is an entirely new point of view in the theory of knot and link invariants. The new picture provides a geometrical interpretation of the integer coefficients of the quantum group invariants, an issue that has been investigated during many years. To present an account of these developments, one needs to review first some basic facts of large- N expansions.

To consider the presence of Wilson loops, it is convenient to introduce a particular generating functional. First, one performs a change of basis from representations R to conjugacy classes $C(k)$ of the symmetric group, labeled by vectors $k = (k_1, k_2, \dots)$ with $k_i \geq 0$, and $|k| = \sum_j k_j > 0$. The change of basis is $W_k = \sum_R \chi_R(C(k)) W_R$, where χ_R are characters of the permutation group S_ℓ of $\ell = \sum_j j k_j$ elements (ℓ is also the number of boxes of the Young tableau associated to R). Second, one introduces the generating functional:

$$F(V) = \log Z(V) = \sum_k \frac{|C(k)|}{\ell!} W_k^{(c)} \Upsilon_k(V) \quad [53]$$

where

$$Z(V) = \sum_k \frac{|C(k)|}{\ell!} W_k \Upsilon_k(V) \\ \Upsilon_k(V) = \prod_j (\text{tr } V^j)^{k_j}$$

In these expressions $|C(k)|$ denotes the number of elements of the class $C(k)$ in S_ℓ . The reason behind the introduction of this generating functional is that the large- N structure of the connected Wilson loops, $W_k^{(c)}$, turns out to be very simple:

$$\frac{|C(k)|}{\ell!} W_k^{(c)} = \sum_{g=0}^{\infty} x^{2g-2+|k|} F_{g,k}(\lambda) \quad [54]$$

where $\lambda = e^t$ and $t = Nx$ is the 't Hooft coupling. Writing $x = t/N$, it corresponds to a power series expansion in $1/N$. As before, the expansion looks like a perturbative series in string theory where g is the genus and $|k|$ is the number of holes. Ooguri and Vafa conjectured in 1999 the appropriate string-theory description of [54]. It corresponds to an open topological string theory (notice that the ones

described in the previous section were closed), whose target space is the resolved conifold X . The contribution from this theory will lead to open-string analogs of Gromov–Witten invariants.

In order to describe in more detail the fact that one is dealing with open strings, some new data need to be introduced. Here is where the knot description intrinsic to the Wilson loop enters. Given a knot K on S^3 , let us associate to it a Lagrangian submanifold C_K with $b_1 = 1$ in the resolved conifold X and consider a topological open string on it. The contributions in this open topological string are localized on holomorphic maps $f: \Sigma_{g,h} \rightarrow X$ with $h = |k|$ which satisfy $f_*[\Sigma_{g,h}] = Q$, and $f_*[C] = j[\gamma]$ for k_j oriented circles C . In these expressions $\gamma \in H_1(C_K, \mathbb{Z})$, and $Q \in H_2(X, C_K, \mathbb{Z})$, that is, the map is such that k_j boundaries of $\Sigma_{g,h}$ wrap the knot j times, and $\Sigma_{g,h}$ itself gets mapped to a relative two-homology class characterized by the Lagrangian submanifold C_K . The number of such maps (in the sense described in the previous section) is the open-string analog of Gromov–Witten invariants. They will be denoted by $N_{g,k}^Q$. Comparing to the situation that led to [47] in the closed-string case, one concludes that in this case the quantities $F_{g,k}(\lambda)$ in [54] must take the form

$$F_{g,k}(\lambda) = \sum_Q N_{g,k}^Q e^{\int_Q \omega}, \quad t = \int_{P^1} \omega \quad [55]$$

where ω is the Kähler class of the Calabi–Yau manifold X and $\lambda = e^t$. For any Q , one can always write $\int_Q \omega = Qt$, where Q is in general a half-integer number. Therefore, $F_{g,k}(\lambda)$ is a polynomial in $\lambda^{\pm 1/2}$ with rational coefficients.

The result [55] is very impressive but still does not provide a representation where one can assign a geometrical interpretation to the integer coefficients of the quantum-group invariants. Notice that to match a polynomial invariant to [55], after obtaining its connected part, one must expand it in x after setting $q = e^x$ keeping λ fixed. One would like to have a refined version of [55], in the spirit of what was described in the previous section leading from the Gromov–Witten invariants N_g^β of [47] to the new integer invariants n_g^β of [48]. It turns out that, indeed, $F(V)$ can be expressed in terms of integer invariants in complete analogy with the description presented in the previous section for topological strings. A good review on the subject can be found in Mariño (2005).

Concluding Remarks

In this overview we have introduced key features of TQFTs and we have described some of the most relevant results emerged from them. We have

described how the many faces of TQFT provide a variety of important insights in a selected set of problems in topology. Among these outstand the reformulation of Donaldson theory and the discovery of the Seiberg–Witten invariants, and the string-theory description of the large- N expansion of Chern–Simons gauge theory, which provides an entirely new point of view in the study of knot and link invariants and points to an underlying fascinating interplay between string theory, knot theory, and enumerative geometry which opens new fields of study.

In addition to their intrinsic mathematical interest, TQFTs have been found relevant to important questions in physics as well. This is so because, in a sense, TQFTs are easier to solve than conventional quantum field theories. For example, topological sigma models are relevant to the computation of certain couplings in string theory. Also, Witten-type gauge TQFTs such as Donaldson–Witten theories and its generalizations play a role in string theory as effective world-volume theories of extended string states (branes) wrapping curved spaces, and TQFTs arising from $\mathcal{N} = 4$ gauge theories in four dimensions have shed light on field- (and string-) theory dualities.

Most of these developments, and others that we have not touched upon or only mentioned in passing have their own entries in the encyclopedia, to which we refer the interested reader for further details.

See also: Axiomatic Approach to Topological Quantum Field Theory; BF Theories; Chern–Simons Models; Rigorous Results; Donaldson–Witten Theory; Gauge Theoretic Invariants of 4-Manifolds; Gauge Theory: Mathematical Applications; Hamiltonian Fluid Dynamics; The Jones Polynomial; Knot Theory and Physics; Mathai–Quillen Formalism; Mathematical Knot Theory; Schwarz-Type Topological Quantum Field Theory; Seiberg–Witten Theory; Stationary Phase Approximation; Topological Sigma Models.

Further Reading

- Atiyah MF (1988) New invariants of three and four dimensional manifolds. In: *The Mathematical Heritage of Herman Weyl*, Proc. Symp. Pure Math., vol. 48. American Math. Soc. pp. 285–299.
- Birmingham D, Blau M, Rakowski M, and Thompson G (1991) Topological field theory. *Physics Reports* 209: 129–340.
- Cordes S, Moore G, and Rangoolam S (1996) Lectures on 2D Yang–Mills theory, equivariant cohomology and topological field theories. In: David F, Ginsparg P, and Zinn-Justin J (eds.) *Fluctuating Geometries in Statistical Mechanics and Field Theory*, Les Houches Session LXII, p. 505 (hep-th/9411210). Elsevier.
- Donaldson SK (1990) Polynomial invariants for smooth four-manifolds. *Topology* 29: 257–315.
- Floer A (1987) Morse theory for fixed points of symplectic diffeomorphisms. *Bulletin of the American Mathematical Society* 16: 279.

Freyd P, Yetter D, Hoste J, Lickorish WBR, Millet K, and Ocneanu A (1985) A new polynomial invariant of knots and links. *Bulletin of the American Mathematical Society* 12(2): 239–246.

Gopakumar R and Vafa C (1999) On the gauge theory/geometry correspondence. *Advances in Theoretical and Mathematical Physics* 3: 1415 (hep-th/9811131).

Jones VFR (1985) A polynomial invariant for knots via von Neumann algebras. *Bulletin of the American Mathematical Society* 12: 103–112.

Jones VFR (1987) Hecke algebra representations of braid groups and link polynomials. *Annals of Mathematics* 126(2): 335–388.

Kauffman LH (1990) An invariant of regular isotopy. *Transactions of American Mathematical Society* 318(2): 417–471.

Labastida JMF (1999) Chern-Simons gauge theory: ten years after. In: Falomir H, Gamboa R, and Schaposnik F (eds.) *Trends in Theoretical Physics II*, ch. 484, 1. New York: AIP (hep-th/9905057).

Labastida JMF and Lozano C (1998) Lectures in topological quantum field theory. In: Falomir H, Gamboa R, and Schaposnik F (eds.) *Trends in Theoretical Physics*, ch. 419, 54. New York: AIP (hep-th/9709192).

Labastida JMF and Mariño M (2005) *Topological Quantum Field Theory and Four Manifolds*. Dordrecht: Elsevier; Norwell, MA: Springer.

Mariño M (2005) Chern-Simons theory and topological strings. *Reviews of Modern Physics* 77: 675–720.

Moore G and Witten E (1998) Integrating over the u -plane in Donaldson theory. *Advances in Theoretical and Mathematical Physics* 1: 298–387.

Vassiliev VA (1990) Cohomology of knot spaces. In: *Theory of Singularities and Its Applications*, Advances in Soviet Mathematics, vol. 1, pp. 23–69. American Mathematical Society.

Witten E (1982) Supersymmetry and Morse theory. *Journal of Differential Geometry* 17: 661–692.

Witten E (1988a) Topological quantum field theory. *Communications in Mathematical Physics* 117: 353.

Witten E (1988b) Topological sigma models. *Communications in Mathematical Physics* 118: 411.

Witten E (1989) Quantum field theory and the Jones polynomial. *Communications in Mathematical Physics* 121: 351.

Witten E (1994) Monopoles and four-manifolds. *Mathematical Research Letters* 1: 769–796.

Topological Sigma Models

D Birmingham, University of the Pacific, Stockton, CA, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Topological sigma models govern the quantum mechanics of maps from a Riemann surface Σ to a target space M . In contrast to the standard supersymmetric sigma model, the topological version has a special local shift symmetry. This symmetry takes the form $\delta u^i = \epsilon^i$, where ϵ^i is an arbitrary local function of the coordinates on the base manifold Σ . In essence, this topological shift symmetry ensures that all local degrees of freedom of the model can be gauged away. As a result, the dynamics of such a model resides in a finite number of global topological degrees of freedom. This feature is generic to all topological field theories of Witten type, also known as cohomological field theories (see Topological Quantum Field Theory: Overview). The topological shift symmetry is responsible for the special topological nature of the model, which is seen most readily by BRST quantizing the local shift symmetry. This gives rise to a nilpotent BRST operator Q . The properties of this BRST operator are crucial for establishing the topological nature of the model. The key point in the construction of any cohomological field theory is the fact that the full quantum action S_q can be written as a BRST commutator $S_q = \{Q, V\}$, where V is a function of the

fields needed to define the path integral. In particular, one can show that the partition function and all correlation functions are independent of the metric on both the base manifold Σ and the target space M . For example, let us define the path integral by

$$Z = \int d\Phi e^{-\{Q,V\}} \tag{1}$$

where Φ denotes the full set of fields required at the quantum level. In general, the function V depends on geometric data of both Σ and M . Nevertheless, one can easily establish that the partition function is independent of this data by noting the following. Variation of Z with respect to the metric of the target space g (for example) gives

$$\delta_g Z = - \int d\Phi e^{-\{Q,V\}} \{Q, \delta_g V\} \tag{2}$$

The right-hand side of this equation is nothing but the vacuum expectation value of a BRST commutator, and this vanishes by BRST invariance of the vacuum. It is important to note here that the BRST operator Q can be constructed to be independent of g . Apart from the necessity of introducing the metric tensor, these models also require additional geometric data for their construction. The complex structure of Σ , and at least an almost-complex structure on M , is required. By a similar argument, one can show that the partition function and correlation functions are independent of this extra

geometric data. As mentioned above, these models possess no local degrees of freedom. One can then show that the path-integral expression for the correlation functions can be localized to a finite-dimensional moduli space of instanton configurations which minimize the classical action.

We will first show how the full quantum action of the theory can be obtained as a BRST quantization of a classical action with a local gauge symmetry. However, we shall then highlight the fact that the gauge algebra for this topological shift symmetry only closes on-shell. In order to proceed with a BRST quantization of the model, and obtain the complete quantum action, one must take recourse to the Batalin–Vilkovisky quantization scheme. This machinery is ideally tailored for such a problem, with the end result that quartic ghost terms are present in the action. However, the presence of such terms does not affect the arguments presented above, since the quantum is still obtained as a BRST commutator. Following this, we construct all observables of the theory and demonstrate their connection to the de Rham cohomology of the target space. The special topological properties of the observables are then discussed, and it is shown how their computation is localized to the moduli space \mathcal{M} of holomorphic maps from Σ to M . As a particular example, we show how the computation of a certain class of observables determines the intersection numbers of the moduli space \mathcal{M} . We present a brief discussion of the connection between topological sigma models with Calabi–Yau target space M , and the mirror symmetry of M .

Construction of the Model

We begin with the following classical action:

$$S_c = \int_{\Sigma} d^2\sigma \sqrt{h} h_{\alpha\beta} g_{ij} K^{\alpha i} K^{\beta j} \quad [3]$$

where

$$K^{\alpha i} = G^{\alpha i} - \frac{1}{2} (\partial^{\alpha} u^i + \epsilon^{\alpha}_{\beta} J^i_j \partial^{\beta} u^j) \quad [4]$$

The fields $G^{\alpha i}$ and $K^{\alpha i}$ both satisfy the self-duality constraint

$$\begin{aligned} G^{\alpha i} &= P^{\alpha i}_{+\beta j} G^{\beta j} \\ K^{\alpha i} &= P^{\alpha i}_{+\beta j} K^{\beta j} \end{aligned} \quad [5]$$

where the self-dual and anti-self-dual projection operators are defined as

$$P^{\alpha i}_{\pm\beta j} = \frac{1}{2} (\delta^{\alpha}_{\beta} \delta^i_j \pm \epsilon^{\alpha}_{\beta} J^i_j) \quad [6]$$

The above action describes a theory of maps $u^i(\sigma)$ from a Riemann surface Σ to an almost complex

manifold M . The coordinates on Σ are denoted by $\sigma^{\alpha} (\alpha=1,2)$, while those on the target manifold M are denoted by $u^i (i=1, \dots, \dim M)$. The metric and complex structure of Σ are denoted by $h_{\alpha\beta}$ and $\epsilon^{\alpha}_{\beta}$, respectively; they obey the relations $\epsilon^{\alpha}_{\beta} \epsilon^{\beta}_{\gamma} = -\delta^{\alpha}_{\gamma}$ and $\epsilon_{\alpha\beta} = h_{\alpha\gamma} \epsilon^{\gamma}_{\beta}$. The metric tensor g_{ij} and almost-complex structure J^i_j of M obey analogous relations to the above. In the general model, the target space need only be an almost-complex manifold. This requires the existence of a globally defined tensor field J^i_j such that $J^i_j J^j_k = -\delta^i_k$.

The action [3] is invariant under the topological shift symmetry

$$\delta u^i = \epsilon^i \quad [7]$$

where ϵ^i is an arbitrary local function of the coordinates on the base manifold Σ . Already, at this level, we see the distinction with the standard sigma model. The presence of this shift symmetry means that all local degrees of freedom can be gauged away, leaving only a finite number of global topological degrees of freedom. It requires some work to determine the corresponding transformation for $G^{\alpha i}$, the key point being the preservation of the self-duality constraint. We find

$$\begin{aligned} \delta G^{\alpha i} &= P^{\alpha i}_{+\beta j} (D^{\beta} \epsilon^j + \frac{1}{2} \epsilon^{\beta}_{\gamma} \epsilon^l (D_l J^j_k) \partial^{\gamma} u^k) \\ &\quad + \frac{1}{2} \epsilon^{\alpha}_{\beta} \epsilon^k (D_k J^i_j) G^{\beta j} - \Gamma^i_{lk} \epsilon^k G^{\alpha l} \end{aligned} \quad [8]$$

where the covariant derivative is defined by $D_{\alpha} \epsilon^i = \partial_{\alpha} \epsilon^i + \Gamma^i_{jk} (\partial_{\alpha} u^j) \epsilon^k$.

Having determined the classical symmetries of the model, we can now proceed with the BRST quantized form of the quantum action. As a topological field theory of Witten type, one can show that the quantum action can be written as a BRST commutator, that is, $S_q = \{Q, V\}$, where the gauge fermion V is defined by

$$V = \int d^2\sigma \sqrt{h} \bar{C}_{\alpha i} \left(\partial^{\alpha} u^i - \frac{\alpha}{4} B^{\alpha i} \right) \quad [9]$$

where α is an arbitrary gauge-fixing parameter. The BRST operator Q is nilpotent $Q^2=0$, off-shell. It is defined by $\delta\Phi = \epsilon\{Q, \Phi\}$, and takes the form

$$\begin{aligned} \delta u^i &= -\epsilon C^i \\ \delta C^i &= 0 \\ \delta \bar{C}_{\alpha i} &= \epsilon \left(B_{\alpha i} + \frac{1}{2} \epsilon^{\alpha}_{\beta} (D_k J^j_i) \bar{C}_{\beta j} C^k + \Gamma^k_{ij} \bar{C}_{\alpha k} C^j \right) \\ \delta B^{\alpha i} &= \frac{\epsilon}{4} C^k C^l (R^i_{kl} + R_{klrs} J^r_s) \bar{C}^{\alpha t} \\ &\quad - \frac{\epsilon}{2} \epsilon^{\alpha}_{\beta} (D_k J^i_j) C^k B^{\beta j} \\ &\quad + \frac{\epsilon}{4} (C^k D_k J^i_s) (C^l D_l J^s_t) \bar{C}^{\alpha t} + \epsilon \Gamma^i_{jk} C^j B^{\alpha k} \end{aligned} \quad [10]$$

In the above, the ghost field is denoted by C^i , while the anti-ghost field $\bar{C}_{\alpha i}$ and the multiplier field $B_{\alpha i}$ obey the self-duality constraint [5]. The key point to note in the above transformations is the fact that the ghost field C^i is BRST invariant. Again, this is a feature which is generic to all cohomological field theories. The existence of such a field allows the construction of an entire set of topological correlation functions, as we shall see in the following section.

While the gauge-fixing parameter α is arbitrary, a conventional choice is to take $\alpha=1$, and then integrate out the multiplier field B . This yields the action in the form

$$\begin{aligned} S_q = \int d^2\sigma \sqrt{h} & \left[\frac{1}{2} h^{\alpha\beta} g_{ij} \partial_\alpha u^i \partial_\beta u^j + \frac{1}{2} \epsilon^{\alpha\beta} J_{ij} \partial_\alpha u^i \partial_\beta u^j \right. \\ & + \bar{C}_{\alpha i} \left(D^\alpha C^i + \frac{1}{2} \epsilon^\alpha_\beta (D_j J^i_k) \partial^\beta u^k C^j \right) \\ & + \frac{1}{8} \bar{C}_\alpha{}^m \bar{C}^{\alpha k} R_{mkjr} C^j C^r \\ & \left. + \frac{1}{16} \bar{C}_{\alpha i} \bar{C}^{\alpha k} (D_j J^{li}) (D_r J_{lk}) C^j C^r \right] \end{aligned} \quad [11]$$

It should be stressed that the classical gauge algebra [7] and [8] only closes on-shell. Quantization of the model is therefore more subtle, and requires use of the Batalin–Vilkovisky formalism. The on-shell closure problem automatically results in the presence of quartic ghost coupling terms in the action and consequently cubic terms in the BRST transformations. Despite this, we have established that the full quantum action can be written as a BRST commutator.

The form of the action simplifies when the complex structure of the target manifold is covariantly constant, $D_k J^i_j = 0$. In this case, the target manifold M is Kähler and we denote the complex coordinates as u^I , with their complex conjugates denoted by $\bar{u}^{\bar{I}}$. The nonzero components of the metric tensor are then $g_{I\bar{J}}$. Similarly, the coordinates of Σ are denoted σ^\pm , with nonzero metric components h_{+-} . The nonzero components of the ghost and anti-ghost are then given by $C^I, \bar{C}^{\bar{I}}, \bar{C}_{+I}, \bar{C}_{-\bar{I}}$. The action can be written in the form

$$\begin{aligned} S_q = \int d^2\sigma \sqrt{h} & \left[h^{+-} g_{I\bar{J}} \partial_+ u^I \partial_- \bar{u}^{\bar{J}} + \frac{1}{2} \bar{C}_{+I} (D_- C^{\bar{I}}) h^{+-} g_{I\bar{J}} \right. \\ & + \frac{1}{2} \bar{C}_{-\bar{I}} (D_+ C^I) h^{+-} g_{I\bar{J}} \\ & \left. + \frac{1}{4} h^{+-} \bar{C}_{+I} \bar{C}_{-\bar{I}} R_{I\bar{J}J\bar{I}} C^I C^{\bar{J}} \right] \end{aligned} \quad [12]$$

Construction of Observables

Having defined the quantum action, it is now of interest to consider the correlation functions of the model. In the functional integral, we integrate over all maps $\Sigma \rightarrow M$ in a fixed homotopy class. Let us consider a correlation function

$$\langle \mathcal{O} \rangle = \int du^i d\bar{C}_{\alpha i} dC^i e^{-tS_q} \mathcal{O} \quad [13]$$

where $t > 0$ is a parameter, and the observable \mathcal{O} is BRST invariant $\{Q, \mathcal{O}\} = 0$. From the BRST invariance of the vacuum, it follows immediately that the vacuum expectation value of a BRST commutator is zero, $\langle \{Q, \mathcal{O}\} \rangle = 0$. An operator which is a BRST commutator is said to be Q -exact. Hence, our interest is in the Q -cohomology classes of operators, that is, BRST invariant operators modulo BRST exact operators. It is for this reason that such a model is called a cohomological field theory.

One can now show that the variation of [13] with respect to t is a BRST commutator, namely

$$\delta_t \langle \mathcal{O} \rangle = -\delta t \int du^i d\bar{C}_{\alpha i} dC^i e^{-tS_q} \{Q, V\mathcal{O}\} = 0 \quad [14]$$

As a result, one can evaluate the correlation function in the large- t (weak-coupling) limit. In this limit, the path integral is dominated by fluctuations around the classical minima. For the sigma model under study, the classical action is minimized by the instanton configurations

$$\partial_\alpha u^i + \epsilon_\alpha{}^\beta J^i_j \partial_\beta u^j = 0 \quad [15]$$

Indeed, this localization of the path integral to the moduli space of instantons can also be seen by choosing the $\alpha=0$ gauge in [9]. Integration over the multiplier field then imposes a delta function constraint to the instanton configurations. The key point in the above derivation is the fact that the quantum action is a BRST commutator, $S_q = \{Q, V\}$. By a similar argument, one can show that variations of $\langle \mathcal{O} \rangle$ with respect to the metric and complex structure of Σ and M are also zero.

Our aim now is to construct the Q -cohomology classes of operators in the theory. Let us first associate an operator $\mathcal{O}_A^{(0)}$ to each p -form $A = A_{i_1 \dots i_p} du^{i_1} \wedge \dots \wedge du^{i_p}$ on the target space M , given by

$$\mathcal{O}_A^{(0)} = A_{i_1 \dots i_p} C^{i_1} \dots C^{i_p} \quad [16]$$

where C^i is the ghost field. Under a BRST transformation, we see that

$$\begin{aligned} \{Q, \mathcal{O}_A^{(0)}\} &= -\partial_{i_0} A_{i_1 \dots i_p} C^{i_0} \dots C^{i_p} \\ &= -\mathcal{O}_{dA}^{(0)} \end{aligned} \quad [17]$$

since the ghost fields are BRST invariant by [10]. Hence, $\mathcal{O}_A^{(0)}$ is BRST invariant if and only if A is a closed p -form. Similarly, if A is an exact p -form, then the corresponding operator is Q -exact. Hence, the BRST cohomology classes of these operators are in one to one correspondence with the de Rham cohomology classes on M . The reason for assigning the peculiar superscript to the operator $\mathcal{O}^{(0)}$ will become clear at the end of this construction. Notice also that operators of the form $\mathcal{O}_A^{(0)}$ can be used as building blocks for constructing new observables. If we consider a set of closed forms A_1, \dots, A_k , then the product of the associated operators $\mathcal{O}_{A_1}^{(0)} \dots \mathcal{O}_{A_k}^{(0)}$ is clearly Q -invariant as well.

When considering the vacuum expectation values of operators which are polynomials in the fields, there is an implicit dependence on the points where the operators are located. In the case at hand however, the operator $\mathcal{O}_A^{(0)}(\sigma)$ at the point σ has a vacuum expectation value which is a topological invariant, and thus cannot depend on the chosen point. To see this explicitly, we consider all fields defined over Σ , and differentiate the operator with respect to some local coordinates σ^α :

$$\begin{aligned} \frac{\partial}{\partial \sigma^\alpha} A_{i_1 \dots i_p} C^{i_1} \dots C^{i_p} &= (\partial_{i_0} A_{i_1 \dots i_p}) \frac{\partial u^{i_0}}{\partial \sigma^\alpha} C^{i_1} \dots C^{i_p} \\ &+ p A_{i_1 \dots i_p} (\partial_{i_0} C^{i_1}) \frac{\partial u^{i_0}}{\partial \sigma^\alpha} C^{i_2} \dots C^{i_p} \end{aligned} \quad [18]$$

In terms of exterior derivatives, this takes the form,

$$\begin{aligned} d\mathcal{O}_A^{(0)} &= \partial_{i_0} A_{i_1 \dots i_p} du^{i_0} C^{i_1} \dots C^{i_p} + p A_{i_1 \dots i_p} dC^{i_1} C^{i_2} \dots C^{i_p} \\ &= \{Q, \mathcal{O}_A^{(1)}\} \end{aligned} \quad [19]$$

where $\mathcal{O}_A^{(1)} = -p A_{i_1 \dots i_p} du^{i_1} C^{i_2} \dots C^{i_p}$, and we have used the fact that A is a closed p -form. If we let γ represent any path between two arbitrary points P and P' , then this expression has the integral form,

$$\mathcal{O}_A^{(0)}(P) - \mathcal{O}_A^{(0)}(P') = \left\{ Q, \int_\gamma \mathcal{O}_A^{(1)} \right\} \quad [20]$$

and we see that the vacuum expectation value of $\mathcal{O}_A^{(0)}$ is point independent by the BRST invariance of the vacuum. The same remark applies to any product of operators of the form we are considering.

To continue our construction, consider a one-dimensional homology cycle $\gamma(\partial\gamma=0)$, and define

$$W_A^{(1)}(\gamma) = \int_\gamma \mathcal{O}_A^{(1)} \quad [21]$$

This new operator $W_A^{(1)}(\gamma)$ is BRST invariant by inspection,

$$\{Q, W_A^{(1)}(\gamma)\} = \int_\gamma \{Q, \mathcal{O}_A^{(1)}\} = \int_\gamma d\mathcal{O}_A^{(0)} = 0 \quad [22]$$

Moreover, if γ happens to be the boundary of a two-dimensional surface ($\gamma = \partial\beta$), so that γ is trivial in homology, then this new operator is likewise trivial in Q cohomology:

$$W_A^{(1)}(\gamma) = \int_\gamma \mathcal{O}_A^{(1)} = \int_\beta d\mathcal{O}_A^{(0)} = \left\{ Q, \int_\beta \mathcal{O}_A^{(2)} \right\} \quad [23]$$

where

$$\mathcal{O}_A^{(2)} = -\frac{p(p-1)}{2} A_{i_1 \dots i_p} du^{i_1} \wedge du^{i_2} C^{i_3} \dots C^{i_p}$$

As before, let us now associate to each homology 2-cycle $\beta(\partial\beta=0)$, another BRST invariant operator $W_A^{(2)}$ defined by

$$W_A^{(2)}(\beta) = \int_\beta \mathcal{O}_A^{(2)} \quad [24]$$

The BRST invariance follows trivially as in [23].

In summary, we have produced three operators $\mathcal{O}_A^{(0)}$, $\mathcal{O}_A^{(1)}$, and $\mathcal{O}_A^{(2)}$ from any given closed form A , which satisfy the relations:

$$\begin{aligned} 0 &= \{Q, \mathcal{O}_A^{(0)}\}, & d\mathcal{O}_A^{(0)} &= \{Q, \mathcal{O}_A^{(1)}\} \\ d\mathcal{O}_A^{(1)} &= \{Q, \mathcal{O}_A^{(2)}\}, & d\mathcal{O}_A^{(2)} &= 0 \end{aligned} \quad [25]$$

The BRST observables are then given by arbitrary products of the integrated operators $W_A^{(i)}(\gamma) = \int_\gamma \mathcal{O}_A^{(i)}$, where γ is any i -cycle in homology.

Observables and Intersection Theory

Let us consider the computation of the correlation function $\langle \mathcal{O} \rangle$ in the background field method. We first pick a background instanton configuration [15], and then integrate over the quantum fluctuations around that instanton. The relevant part of the quantum action is quadratic in the quantum fields, and localization of the model then ensures that such a computation is exact. The quantum fields are expanded into eigenfunctions of the operators that appear in the quadratic part of the action, and the functional integral is replaced by an integral over the eigenmodes. However, if there are fermionic zero modes, then those modes do not enter in the action. As a result, the fermionic integrals ($\int d\chi = 0$) over those modes will cause $\langle \mathcal{O} \rangle$ to vanish unless it has the correct fermion content; the zero modes must be absorbed. In our case, a glance at the quantum action indicates that we should concern ourselves with the zero modes of the ghost C^i and anti-ghost

$\bar{C}_{\alpha i}$. A C^i zero mode is clearly in the kernel of the operator

$$\bar{D}_{\alpha j}^i = D_{\alpha} \delta^i_j + \epsilon_{\alpha\beta} J^i_j D^{\beta} + \epsilon_{\alpha\beta} (D_j J^i_k) \partial^{\beta} u^k \tag{26}$$

and a $\bar{C}_{\alpha i}$ zero mode is a zero eigenfunction of its adjoint \bar{D}^* . In the BRST quantization of the model, the ghost fields C^i are assigned ghost number +1, while the anti-ghost fields $\bar{C}_{\alpha i}$ have ghost number -1. It is therefore apparent that the vacuum expectation value of any observable will vanish unless that observable has a ghost number equal to the number of \bar{D} zero modes, a , minus the number of \bar{D}^* zero modes, b . This difference, $\omega = a - b$, is called the index of the operator \bar{D} .

There is a direct link between this index and the dimension of the moduli space of instantons. Recall that we are considering the space of maps $\Sigma \rightarrow M$ in a specified homotopy class, which satisfy equation [15]. It is then of interest to determine the dimension of the space of such solutions. To this aim, we examine the constraint that arises by considering an instanton u^i , and another neighboring solution $u^i + \hat{u}^i$, where \hat{u}^i is an infinitesimal deformation. To first order in \hat{u}^i , we see that \hat{u}^i must be a zero mode of the operator \bar{D} . This is no coincidence, and we can thus interpret the ghost fields C^i as cotangent vectors to instanton moduli space \mathcal{M} . In particular, if \mathcal{M} is a smooth manifold, then $\dim \mathcal{M} = a$. The index of the operator \bar{D} is called the virtual dimension of the moduli space. In generic situations, the virtual dimension is equal to the actual dimension $\dim \mathcal{M}$.

It is possible to interpret some of the observables that we have described in terms of intersection theory applied to the moduli space of instantons. In particular, one can show that all correlation functions of the form

$$\langle \mathcal{O}_{A_1}^{(0)} \cdots \mathcal{O}_{A_s}^{(0)} \rangle \tag{27}$$

are intersection numbers of certain submanifolds of moduli space. In order to see this in a simple example, we first recall the notion of Poincaré duality and the relationship between cohomology and homology.

Poincaré duality can be formulated as a relationship between de Rham cohomology (defined in terms of closed differential forms) and homology (defined in terms of subspaces of M). For our purposes here, it is sufficient to state that we can associate to each boundaryless submanifold N of codimension k , a cohomology class $[\phi] \in H^k(M)$, such that

$$\int_M \phi \wedge \psi = \int_N \psi \tag{28}$$

for all $[\psi] \in H^{n-k}(M)$. By ψ on the right-hand side of this equation, we mean the pullback $i^* \psi$ under the inclusion $i: N \rightarrow M$. Conversely, to each closed k -form ϕ on M , we can associate an $(n - k)$ -cycle N (it is in general a chain of subspaces), unique up to homology, such that the previous relation is satisfied. Furthermore, one can show that the Poincaré dual to N can be chosen in such a way that its support is localized within any given open neighborhood of N in M (essentially delta function support on N).

Let us now define the notion of transversal intersection. For simplicity, we will first consider the intersection of two submanifolds M_1 and M_2 contained in M . We will say that these two submanifolds have transversal intersection if the tangent spaces satisfy

$$T_x(M_1) + T_x(M_2) = T_x(M) \tag{29}$$

for all $x \in M_1 \cap M_2$. It is a theorem that a submanifold of codimension k can be locally “cut-out” by k smooth functions, that is, the submanifold is locally specified by the zeros of this set of functions. It is a worthwhile exercise to convince oneself that the definition of transversal intersection is equivalent to the statement that the functions which cut-out M_1 are independent from those which cut-out M_2 . Thus, we can write

$$\text{codim}(M_1 \cap M_2) = \text{codim}(M_1) + \text{codim}(M_2) \tag{30}$$

More generally, we say that the intersection $M_1 \cap \cdots \cap M_s$ of s submanifolds is transversal if the intersection of every pair of them is transversal. It then follows trivially by the previous argument that the codimensions must satisfy

$$\text{codim}(M_1 \cap \cdots \cap M_s) = \sum_{i=1}^s \text{codim}(M_i) \tag{31}$$

The special case which will be important for us occurs when the intersection of submanifolds is a collection of points, that is, when the codimension of the intersection is equal to the dimension of M . Since these points are isolated, the compactness of M guarantees that they are finite in number.

We are now in a position to describe in what sense correlation functions of the form $\langle \mathcal{O}_{A_1}^{(0)} \cdots \mathcal{O}_{A_s}^{(0)} \rangle$ determine intersection numbers in the moduli space \mathcal{M} of instantons. By definition, this moduli space is the set of maps from Σ to M which satisfy [15]. Let us consider the generic situation, where the virtual dimension of \mathcal{M} (i.e., the index of \bar{D}) is equal to $\dim \mathcal{M}$. For convenience, let us begin by choosing the forms A_i which represent de Rham cohomology classes on M , together with their Poincaré duals M_i , such that the forms have essentially delta function

support on their respective submanifolds. Since each of the operators in the correlation function depends on some fixed point σ_i , it is meaningful to define the submanifolds $L_i \equiv \{u \in \mathcal{M} \mid u(\sigma_i) \in M_i\} \subset \mathcal{M}$. Now, the correlation function represents a functional integral over the space of maps $\text{Map}(\Sigma, M)$, and we have argued that this integral only receives contributions from the instanton configurations. Since the operators $A_i(u(\sigma_i))$ vanish unless $u \in L_i$ by our choice of the Poincaré duals, we see that the only contribution to the functional integral can be from those maps which lie in the intersection $L_1 \cap \dots \cap L_s$. By ghost number considerations, this correlation function must vanish unless the codimension of the intersection equals the virtual dimension of \mathcal{M} . In the generic case where the virtual dimension is equal to $\dim \mathcal{M}$, this means that the intersection is simply a finite number of points. Intersection numbers ± 1 can then be assigned to each point in the intersection $L_1 \cap \dots \cap L_s$, by considering the relative orientation of the submanifolds L_i at the intersection points. From the functional integral point of view, the computation reduces to an evaluation of the ratio of the bosonic determinant (integration over u^i) to the fermionic determinant (integration over C^i and $\bar{C}_{\alpha i}$). In the Kähler case, for example, the intersection number assigned to each point in the intersection is always $+1$. This is due to the fact that the C^i, \bar{C}_{-}^i determinant is the complex conjugate of the C^i, \bar{C}_{+}^i determinant.

A and B Models and Mirror Symmetry

The topological sigma model for a Kähler target space [12] is also known as the topological A model. In this case, the action can be recovered by twisting the standard $N=2$ supersymmetric sigma model. This twisting procedure amounts to a reassignment of the spins of the fields in the theory. However, there is an alternative twisting which can be done, and this leads to another model known as the

topological B model. The usefulness of this observation lies in the fact that the topological A model on a Calabi–Yau target space M is related to the topological B model on the mirror of M . This relationship and the computation of correlation functions in the A and B models thus sheds light on the nature of mirror symmetry.

See also: Batalin–Vilkovisky Quantization; BRST Quantization; Functional Integration in Quantum Physics; Graded Poisson Algebras; Mathai–Quillen Formalism; Mirror Symmetry: A Geometric Survey; Several Complex Variables: Compact Manifolds; Singularities of the Ricci Flow; Topological Gravity, Two-Dimensional; Topological Quantum Field Theory: Overview; WDVV Equations and Frobenius Manifolds.

Further Reading

- Baulieu L and Singer I (1989) The topological sigma model. *Communications in Mathematical Physics* 125: 227.
- Birmingham D, Blau M, Rakowski M, and Thompson G (1991) Topological field theory. *Physics Reports* 209: 129.
- Birmingham D, Rakowski M, and Thompson G (1989) BRST quantization of topological field theory. *Nuclear Physics B* 315: 577.
- Eguchi T and Yang SK (1990) $N=2$ superconformal models as topological field theories. *Modern Physics Letters A* 5: 1693.
- Floer A (1987) Morse theory for fixed points of symplectic diffeomorphisms. *Bulletin of the American Mathematical Society* 16: 279.
- Floer A (1988) An instanton invariant for three manifolds. *Communications in Mathematical Physics* 118: 215.
- Gromov M (1985) Pseudo holomorphic curves in symplectic manifolds. *Inventiones Mathematicae* 82: 307.
- Gromov M (1986) In: *Proceedings of the International Congress of Mathematicians*, p. 81. Berkeley.
- Vafa C (1991) Topological Landau–Ginzburg models. *Modern Physics Letters A* 6: 337.
- Witten E (1988) Topological sigma models. *Communications in Mathematical Physics* 118: 411.
- Witten E (1991) Mirror manifolds and topological field theory. In: Yau ST (ed.) *Mirror Symmetry I*, p. 121. American Mathematical Society.

Turbulence Theories

R M S Rosa, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Turbulence has initially been defined as an irregular motion in fluids. The cloud formations in the atmosphere and the motion of water in rivers make this point clear. These are but a few readily available

examples of a multitude of flows which display turbulent regimes: from the blood that flows in our veins and arteries to the motion of air within our lungs and around us; from the flow of water in creeks to the atmospheric and oceanic currents; from the flows past submarines, ships, automobiles, and aircraft to the combustion processes propelling them; and in the flow of gas, oil, and water, from the prospecting end to the entrails of the cities. The great majority of flows in nature and in engineering applications are somehow turbulent.

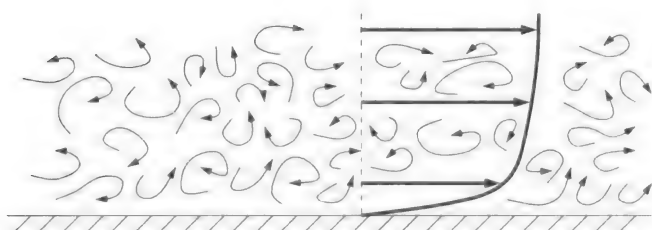


Figure 1 Illustration of the irregular motion of a turbulent flow over a flat plate (thin lines), and of the well-defined velocity profile of the mean flow (thick lines).

But turbulent flows are much more than simply irregular. More refined definitions were desirable and were later coined. A definitive and precise one, however, may only come when the phenomenon is fully understood. Nevertheless, several characteristic properties of a turbulent flow can be listed:

Irregularity and unpredictability A turbulent flow is irregular both in space and time, displaying unpredictable, random patterns.

Statistical order From the irregularity of a turbulent motion there emerges a certain statistical order. Mean quantities and correlation are regular and predictable (**Figure 1**).

Wide range of active scales A wide range of scales of motion are active and display an irregular motion, yielding a large number of degrees of freedom.

Mixing and enhanced diffusivity The fluid particles undergo complicated and convoluted paths, causing a large mixing of different parts of fluid. This mixing significantly enhances diffusion, increasing the transport of momentum, energy, heat, and other advected quantities.

Vortex stretching When a moving portion of fluid also rotates transversally to its motion an increase in speed causes it to rotate faster, a phenomenon called vortex stretching. This causes that portion of fluid to become thinner and elongated, and fold and intertwine with other such portions. This is an intrinsically three-dimensional mechanism which plays a fundamental role in turbulence and is associated with large fluctuations in the vorticity field.

Turbulent Regimes

Turbulence is studied from many perspectives. The subject of “transition to turbulence” attempts to describe the initial mechanisms responsible for the generation of turbulence starting from a laminar motion in particular geometries. This transition can be followed with respect to position in space (e.g., the flow becomes more complicated as we look further downstream on a flow past an obstacle or

over a flat plate) or to parameters (e.g., as we increase the angle of attack of a wing or the pressure gradient in a pipe). This subject is divided into two cases: wall-bounded and free-shear flows. In the former, the viscosity, which causes the fluid to adhere to the surface of the wall, is the primary cause of the instability in the transition process. In the latter, inviscid mechanisms such as mixing layers and jets are the main factors. The tools for studying the transition to turbulence include linearization of the equations of motion around the laminar solution, nonlinear amplitude equations, and bifurcation theory.

“Fully developed turbulence,” on the other hand, concerns turbulence which evolves without imposed constraints, such as boundaries and external forces. This can be thought of turbulence in its “pure” form, and it is somewhat a theoretical framework for research due to its idealized nature. Hypotheses of homogeneity (when the mean quantities associated with the statistical order characterizing a turbulent flow are independent in space), stationarity (*idem* in time), and isotropy (*idem* with respect to rotations in space) concern fully developed turbulent flows. The Kolmogorov theory was developed in this context and it is the most fundamental theory of turbulence. Current research is dedicated in great part to unveil the mechanisms behind a phenomenon called intermittency and how it affects the laws obtained from the conventional theory. Research is also dedicated to derive such laws as much from first principles as possible, minimizing the use of phenomenological and dimensional analysis.

Real turbulent flows involve various regimes at once. A typical flow past a blunt object, for instance, displays laminar motion at its upstream edge, a turbulent boundary layer further downstream, and the formation of a turbulent wake (**Figure 2**). The subject of turbulent boundary layer is a world in itself with current research aiming to determine mean properties of flows over rough surfaces and varied topography. Convective turbulence involves coupling with active scalars such as



Figure 2 Illustration of a flow past an object, with a laminar boundary layer (light gray), a turbulent boundary layer (medium gray), and a turbulent wake (dark gray).

large heat gradients, occurring in the atmosphere, and large salinity gradients, in the ocean. Geophysical turbulence involves also stratification and the anisotropy generated by Earth's rotation. Anisotropic turbulence is also crucial in astrophysics and plasma theory. Multiphase and multicomponent turbulence appear in flows with suspended particles or bubbles and in mixtures such as gas, water, and oil. Transonic and supersonic flows are also of great importance and fall into the category of compressible turbulence, much less explored than the incompressible case.

In all those real situations one would like, from the engineering point of view, to compute mean properties of the flow, such as drag and lift for more efficient designs of aircraft, ships, and other vehicles. Knowledge of the drag coefficient is also of fundamental importance in the design of pipes and pumps, from pipelines to artificial human organs. Mean turbulent diffusion coefficients of heat and other passive scalars – quantities advected by the flow without interfering on it, such as chemical products, nutrients, moisture, and pollutants – are also of major importance in industry, ecology, meteorology, and climatology, for instance. And in most of those cases a large amount of research is dedicated to the “control of turbulence,” either to increase mixing or reduce drag, for instance. From a theoretical point of view, one would like to fully understand and characterize the mechanisms involved in turbulent flows, clarifying this fascinating phenomenon. This could also improve practical applications and lead to a better control of turbulence.

The concept of “two-dimensional turbulence” is controversial. A two-dimensional flow may be irregular and display mixing, statistical order, and a wide range of active scales but definitely it does not involve vortex stretching since the velocity field is always perpendicular to the vorticity field. For this reason many researchers discard two-dimensional turbulence altogether. It is also argued that real two-dimensional flows are unstable at complicated regimes and soon develop into a three-dimensional flow. Nevertheless, many believe that two-dimensional turbulence, even lacking vortex stretching, is of fundamental theoretical importance. It may shed some light into the three-dimensional theory and modeling, and it can serve as an approximation to some situations such as the motion of the atmosphere and oceans in the large and meso scales and some magnetohydrodynamic flows. The relative shallowness of the atmosphere and oceans or the imposition of a strong uniform magnetic field may force the flow into two-dimensionality, at least for a certain range of scales.

“Chaos” serves as a paradigm for turbulence, in the sense that it is now accepted that turbulence is a dynamic processes in a sensitive deterministic system. But not all chaotic motions in fluids are termed turbulent for they may not display mixing and vortex stretching or involve a wide range of scales. An important such example appears in the dispersive, nonlinear interactions of waves.

The Equations of Motion

It is usually stressed that turbulence is a continuum phenomenon, in the sense that the active scales are much larger than the collision mean free path between molecules. For this reason, turbulence is believed to be fully accounted for by the Navier–Stokes equations.

In the case of incompressible homogeneous flows, the Navier–Stokes equations in the Eulerian form and in vector notation read

$$\frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p = \mathbf{f} \quad [1a]$$

$$\nabla \cdot \mathbf{u} = 0. \quad [1b]$$

Here, $\mathbf{u} = \mathbf{u}(\mathbf{x}, t) = (u_1, u_2, u_3)$ denotes the velocity vector of an idealized fluid particle located at position $\mathbf{x} = (x_1, x_2, x_3)$, at time t . The mass density in a homogeneous flow is constant, denoted ρ . The constant ν denotes the kinematic viscosity of the fluid, which is the molecular viscosity μ divided by ρ . The variable $p = p(\mathbf{x}, t)$ is the kinematic pressure, and $\mathbf{f} = \mathbf{f}(\mathbf{x}, t) = (f_1, f_2, f_3)$ denotes the mass density of volume forces.

Equation [1a] expresses the conservation of linear momentum. The term $\nu \Delta \mathbf{u}$ accounts for the dissipation of energy due to molecular viscosity, and the nonlinear term $(\mathbf{u} \cdot \nabla) \mathbf{u}$, also called the inertial term, accounts for the redistribution of energy among different structures and scales of motion. Equation [1b] represents the incompressibility condition. In Einstein's summation convention, these equations can be written as

$$\frac{\partial u_i}{\partial t} + \nu \frac{\partial^2 u_i}{\partial x_j^2} + u_j \frac{\partial u_i}{\partial x_j} + \frac{\partial p}{\partial x_i} = f_i, \quad \frac{\partial u_j}{\partial x_j} = 0$$

The Reynolds Number

The transition to turbulence was carefully studied by Reynolds in the late nineteenth century in a series of experiments in which water at rest in a tank was allowed to flow through a glass pipe. Starting with dimensional analysis, Reynolds argued that a critical

value of a certain nondimensional quantity was likely to exist beyond which a laminar flow gives rise to a “sinuous” motion. This was followed by observations of the flow for tubes with different diameter L , different mean velocities U across the tube section, and with the kinematic viscosity $\nu = \rho/\mu$ being altered through changes in temperature. The experiments confirmed the existence of such a critical value for what is now called the Reynolds number:

$$Re = \frac{LU}{\nu}$$

The dimensional analysis argument can be reproduced in the following form: the physical dimension for the inertial term in [1a] is U^2/L , while that for the viscous term is $\nu U/L^2$. The ratio between them is precisely $Re = LU/\nu$. For small values of Re viscosity dominates and the flow is laminar, whereas for large values of Re the inertial term dominates, and the flow becomes more complicated and eventually turbulent. In applications, different types of Reynolds number can be used depending on the choice of the characteristic velocity and length, but in any case, the larger the Reynolds number, the more complicated the flow.

The Reynolds Equations

Another advance put forward by Reynolds in a subsequent article was to decompose the flow into a mean component and the remaining fluctuations. In terms of the velocity and pressure fields this can be written as

$$\mathbf{u} = \bar{\mathbf{u}} + \mathbf{u}', \quad p = \bar{p} + p' \quad [2]$$

with $\bar{\mathbf{u}}$ and \bar{p} representing the mean components and \mathbf{u}' and p' , the fluctuations. By substituting [2] into [1], one finds the Reynolds-averaged Navier–Stokes (RANS) equations for the mean flow:

$$\frac{\partial \bar{\mathbf{u}}}{\partial t} - \nu \Delta \bar{\mathbf{u}} + (\bar{\mathbf{u}} \cdot \nabla) \bar{\mathbf{u}} + \nabla \bar{p} = \mathbf{f} + \nabla \cdot \boldsymbol{\tau}$$

$$\nabla \cdot \bar{\mathbf{u}} = 0$$

It differs from [1] only by the addition of the Reynolds stress tensor:

$$\boldsymbol{\tau} = -\overline{\mathbf{u}' \otimes \mathbf{u}'} = -\left(\overline{u'_i u'_j}\right)_{i,j=1}^3$$

In a laminar flow, the fluctuations are negligible, otherwise this decomposition shows how they influence the mean flow through this additional turbulent stresses.

The Closure Problem and Turbulence Models

The RANS equations cannot be solved directly for the mean flow since the Reynolds stresses are unknown. Equations for these stress terms can be derived but they involve further unknown moments. This continues with equations for moments of a given order depending on new moments up to a higher order, leading to an infinite system of equations known as the Friedman–Keller system. For practical applications, approximations closing the system at some finite order are needed, in what is called the closure problem. Several *ad hoc* approximations exist, the most famous being the Boussinesq eddy-viscosity approximation, in which the turbulent fluctuations are regarded as increasing the viscosity of the flow. Prandtl’s mixing-length hypothesis yields a prescription for the computation of this eddy viscosity, and together they form the basis of the algebraic models of turbulence. Other models involve additional equations, such as the k - ϵ and k - ω models. Most of the practical computations of industrial flows are based on such lower-order models, and a large amount of research is done to determine appropriate values for the various *ad hoc* parameters which appear in these models and which are highly dependent on the geometry of the flow. This dependency can be explained by the fact that the RANS is supposed to model the mean flow even at the large scales of motion, which are highly affected by the geometry.

Computational fluid dynamics (CFD) is indeed a fundamental tool in turbulence, both for research and engineering applications. From the theoretical side, direct numerical simulations (DNS), which attempt to resolve all the active scales of the flow, reveal some fundamental mechanisms involved in the transition to turbulence and in vortex stretching. As for applications, DNS applies to flows up to low-Reynolds turbulence, with the current computational power not allowing for a full resolution of all the scales involved in high-Reynolds flows. And the current rate of evolution of computational power predicts that this will continue so for several decades.

An intermediate CFD method between RANS and DNS is the large-eddy simulation (LES), which attempts to fully resolve the large scales while modeling the turbulent motion at the smaller scales. Several models have been proposed which have their own advantages and limitations as compared to RANS and DNS. It is currently a subject of intense research, particularly for the development of suitable models for the structure functions near the boundary. Theoretical results on fully developed turbulence play a fundamental role in the modeling process.

LESs are a promising tool and they have been successfully applied to a number of situations. The choice of the best method for a given application, however, depends very much on the Reynolds number of the flow and the prior knowledge of similar situations for adjusting the parameters.

Elements of the Statistical Theory

Several types of averages can be used. The ensemble average is taken with respect to a number of experiments at nearly identical conditions. Despite the irregular motion of, say, the velocity vector $\mathbf{u}^{(n)}(\mathbf{x}, t)$ of each experiment $n = 1, \dots, N$, the average value

$$\bar{\mathbf{u}}(\mathbf{x}, t) = \frac{1}{N} \sum_{n=1}^N \mathbf{u}^{(n)}(\mathbf{x}, t)$$

is expected to behave in a more regular way. This type of averaging is usually denoted with the symbol $\langle \cdot \rangle$. This notion can be cast into the context of a probability space $(\mathcal{M}, \Sigma, \mathcal{P})$, where \mathcal{M} is a set, Σ is a σ -algebra of subsets of \mathcal{M} , and \mathcal{P} is a probability measure on Σ . The velocity field is a random variable in the sense that it is a density function $\omega \mapsto \mathbf{u}(\mathbf{x}, t, \omega)$ from \mathcal{M} into the space of time-dependent divergence-free velocity fields. The mean velocity field in this context is regarded as

$$\langle \mathbf{u}(\mathbf{x}, t) \rangle = \int_{\mathcal{M}} \mathbf{u}(\mathbf{x}, t, \omega) d\mathcal{P}(\omega)$$

Other flow quantities such as energy and correlations in space and time can be expressed by means of a function $\varphi = \varphi(\mathbf{u}(\cdot, \cdot))$ of the velocity field, with their mean value given by

$$\langle \varphi(\mathbf{u}(\cdot, \cdot)) \rangle = \int_{\mathcal{M}} \varphi(\mathbf{u}(\cdot, \cdot, \omega)) d\mathcal{P}(\omega)$$

In general, the statistics of the flow are allowed to change with time. A particular situation is when statistical equilibrium is reached, so that $\langle \mathbf{u}(\mathbf{x}, t) \rangle$ and, more generally, $\langle \varphi(\mathbf{u}(\cdot, \cdot + t)) \rangle$ are independent of t . In this case, an ergodic assumption is usually invoked, which means that for “most” individual flows $\mathbf{u}(\cdot, \cdot, \omega_0)$ (i.e., for almost all ω_0 with respect to the probability measure \mathcal{P}), the time averages along this flow converge to the mean ensemble value as the period of the average increases to the mean value obtained by the ensemble average:

$$\begin{aligned} \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi(\mathbf{u}(\cdot, \cdot + s, \omega_0)) ds \\ = \int_{\mathcal{M}} \varphi(\mathbf{u}(\cdot, \cdot, \omega)) d\mathcal{P}(\omega) \end{aligned}$$

Based on this assumption, the averages may in practice be calculated as time averages over a sufficiently large period T . There is a related argument for substituting space averages by time averages and based on the mechanics of turbulence which is called the “Taylor hypothesis.”

Another fundamental concept in the statistical theory is that of homogeneity, which is the spatial analog of the statistical equilibrium in time. In homogeneous turbulence, the statistical quantities of a flow are independent of translations in space, that is,

$$\langle \varphi(\mathbf{u}(\cdot + \ell, \cdot)) \rangle = \langle \varphi(\mathbf{u}(\cdot, \cdot)) \rangle$$

for all $\ell \in \mathbb{R}^3$. The concept of isotropic turbulence assumes further independence with respect to rotations and reflections in the frame of reference, that is,

$$\langle \varphi(Q^t \mathbf{u}(Q \cdot, \cdot)) \rangle = \langle \varphi(\mathbf{u}(\cdot, \cdot)) \rangle$$

for all orthogonal transformations Q in \mathbb{R}^3 , with adjoint Q^t .

Under the homogeneity assumption, mean quantities can be defined independently of position in space, such as the mean kinetic energy per unit mass

$$e = \frac{1}{2} \langle |\mathbf{u}(\mathbf{x})|^2 \rangle = \frac{1}{2} \sum_{i=1}^3 \langle |u_i(\mathbf{x})|^2 \rangle$$

and the mean rate of viscous energy dissipation per unit mass and unit time

$$\epsilon = \nu \sum_{i=1}^3 \langle |\nabla u_i(\mathbf{x})|^2 \rangle = \nu \sum_{i,j=1}^3 \left\langle \left| \frac{\partial u_i(\mathbf{x})}{\partial x_j} \right|^2 \right\rangle$$

The mean kinetic energy can be written as $e = \text{tr}R(0)/2$, where

$$\text{tr}R(\ell) = R_{11}(\ell) + R_{22}(\ell) + R_{33}(\ell), \quad \ell \in \mathbb{R}^3,$$

is the trace of the correlation tensor

$$\begin{aligned} R(\ell) &= \langle \mathbf{u}(\mathbf{x}) \otimes \mathbf{u}(\mathbf{x} + \ell) \rangle = (R_{ij}(\ell))_{i,j=1}^3 \\ &= (\langle u_i(\mathbf{x}) u_j(\mathbf{x} + \ell) \rangle)_{i,j=1}^3 \end{aligned}$$

which measures the correlation between the velocity components at different positions in space. From the homogeneity assumption, this tensor is a function only of the relative position ℓ . Then, assuming that the Fourier transform of $\text{tr}R(\ell)$ exists, and denoting it by $Q(\kappa)$, for $\kappa \in \mathbb{R}^3$, we have

$$\begin{aligned} \text{tr}R(\ell) &= \frac{1}{(2\pi)^{3/2}} \int_{\mathbb{R}^3} Q(\kappa) e^{i\ell \cdot \kappa} d\kappa \\ &= 2 \int_0^\infty S(\kappa) e^{i\ell \cdot \kappa} d\kappa \end{aligned}$$

where $\mathcal{S}(\kappa)$ is the energy spectrum defined by

$$\mathcal{S}(\kappa) = \frac{1}{2(2\pi)^{3/2}} \int_{|\kappa|=\kappa} Q(\kappa) d\Sigma(\kappa) \\ \forall \kappa > 0$$

with $d\Sigma(\kappa)$ denoting the area element of the 2-sphere of radius $|\kappa|$. Then we can write

$$e = \frac{1}{2} \langle |\mathbf{u}(\mathbf{x})|^2 \rangle = \frac{1}{2} \text{tr} R(0) \\ = \int_0^\infty \mathcal{S}(\kappa) d\kappa$$

By expanding the velocity coordinates into Fourier modes $\exp(\ell \cdot \kappa)$, with $\kappa \leq |\kappa| \leq \kappa + d\kappa$ and interpreting them as “eddies” with characteristic wave number $|\kappa|$, the quantity $\mathcal{S}(\kappa)d\kappa$ can be interpreted as the energy of the component of the flow formed by the “eddies” with characteristic wave number between κ and $\kappa + d\kappa$.

Similarly,

$$\epsilon = 2\nu \int_0^\infty \kappa^2 \mathcal{S}(\kappa) d\kappa$$

and we obtain the dissipation spectrum $2\nu\kappa^2\mathcal{S}(\kappa)$, which can be interpreted as the density of energy dissipation occurring at wave number κ .

In the previous arguments it is assumed that the flow extends to all the space \mathbb{R}^3 . This avoids the presence of boundaries, addressing the idealized case of fully developed turbulence. It is sometimes customary to assume as well that the flow is periodic in space to avoid problems with unbounded domains such as infinite kinetic energy.

The random nature of turbulent flows was greatly explored by Taylor in the early twentieth century, who introduced most of the concepts described above. Another important concept he introduced was the Taylor microlength ℓ_T , which is a characteristic length for the small scales based on the correlation tensor. A microscale Reynolds number based on the Taylor microlength is very often used in applications.

Kolmogorov Theory

An inspiring concept in the theory of turbulence is Richardson’s “energy cascade” process. For large Reynolds numbers the nonlinear term dominates the viscosity according to the dimensional analysis, but this is valid only for the large-scale structures. The small scales have their own characteristic length and velocity. In the cascade process, the inertial term is responsible for the transfer of energy to smaller and smaller scales until small enough scales are reached

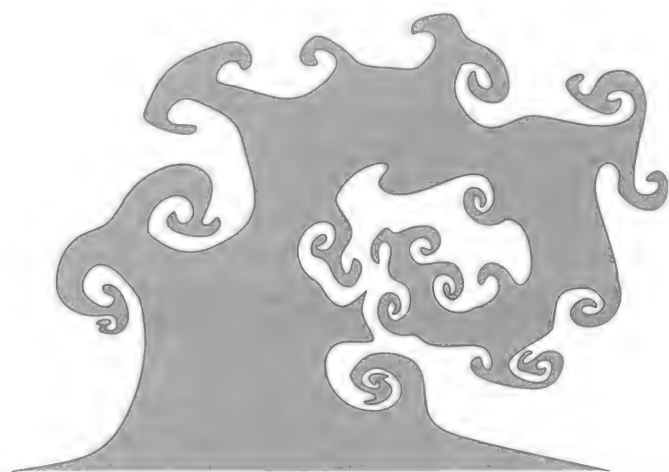


Figure 3 Illustration of the eddy breakdown process in which energy is transferred to smaller eddies and so on until the smallest scales are reached and the energy is dissipated by viscosity.

for which viscosity becomes important (Figure 3). At those smallest scales kinetic energy is finally dissipated into heat. It should be emphasized that turbulence is a dissipative process; no matter how large the Reynolds number is, viscosity plays a role in the smallest scales.

The Kolmogorov theory of locally isotropic turbulence allows for inhomogeneity and anisotropy in the large scales, which contain most of the energy, assuming that with the cascade transfer of energy to smaller scales, the orienting effects generated in the large scales become weaker and weaker so that for sufficiently small eddies the motion becomes statistically homogeneous, isotropic, and independent of the particular energy-productive mechanisms. He proposed that the statistical regime of the small-scale eddies is then universal and depends only on ν and ϵ . The equilibrium range is defined as the range of scales in which this universality holds.

Simple dimensional analysis shows that the only algebraic combination of ν and ϵ with dimension of length is $\ell_\epsilon = (\nu^3/\epsilon)^{1/4}$, which is then interpreted as that near which the viscous effect becomes important and hence most of the energy dissipation takes place. The scale ℓ_ϵ is known as Kolmogorov dissipation length.

Kolmogorov theory gives particular attention to moments involving differences of velocities, such as the p th-order structure function

$$S_p(\ell) \stackrel{\text{def}}{=} \langle (\mathbf{u}(\mathbf{x} + \ell \mathbf{e}) \cdot \mathbf{e} - \mathbf{u}(\mathbf{x}) \cdot \mathbf{e})^p \rangle$$

where \mathbf{e} may be taken as an arbitrary unit vector, thanks to the isotropy assumption. By restricting the search for universal laws for the structure functions only for small values of ℓ anisotropy and inhomogeneity are allowed in the large scales.

The theory assumes a wide separation between the energy-containing scales, of order say ℓ_0 , and the energy-dissipative scales, of order ℓ_ϵ , so that the cascade process occurs within a wide range of scales ℓ such that $\ell_0 \gg \ell \gg \ell_\epsilon$. In this range, termed the inertial range, the viscous effects are still negligible and the statistical regime should depend only on ϵ . Then, the Kolmogorov “two-thirds law” asserts that within the inertial range the second-order correlations must be proportional to $(\epsilon \ell)^{2/3}$, that is,

$$S_2(\ell) = C_K(\epsilon \ell)^{2/3}$$

for some constant C_K known as the Kolmogorov constant in physical space (there is a related constant in spectral space). The argument extends to higher-order structure functions, yielding

$$S_p(\ell) = C_p(\epsilon \ell)^{p/3}$$

Kolmogorov’s derivation of these results was not by dimensional analysis, it was in fact a more convincing self-similarity argument based on the universality assumed for the equilibrium range. A different argument without resorting to universality assumptions, however, was applied to the third-order structure function, yielding the more precise “four-fifths law”:

$$S_3(\ell) = -\frac{4}{5}\epsilon \ell$$

The “Kolmogorov five-thirds law” concerns the energy spectrum $S(\kappa)$ and is the spectral version of the two-thirds law, given by Obukhoff:

$$S(\kappa) = C'_K \epsilon^{2/3} \kappa^{-5/3}$$

The constant C'_K is the Kolmogorov constant in spectral space. The spectral version of the dissipation length is the Kolmogorov wave number $\kappa_\epsilon = (\epsilon/\nu^3)^{1/4}$.

A typical distribution of energy in a turbulent flow is depicted in Figure 4. The energy is

concentrated on the large scales, while the dissipation is concentrated near the Kolmogorov scale ℓ_ϵ . The four-fifths law becomes visible as a straight line in the logarithmic scale.

A more precise mechanism for the energy cascade assumes that in the inertial range, eddies with length scale ℓ transfer kinetic energy to smaller eddies during their characteristic timescale, also known as circulation time. If u_ℓ is their characteristic velocity, then $\tau_\ell = \ell/u_\ell$ is their circulation time, so that the kinetic energy transferred from these eddies during this time is

$$\epsilon_\ell \sim \frac{u_\ell^2}{\tau_\ell} = \frac{u_\ell^3}{\ell}$$

In statistical equilibrium, the energy lost to the smaller scales equals the energy gained from the larger scales, and that should also equal the total kinetic energy dissipated by viscous effects. Hence, $\epsilon_\ell \equiv \epsilon$, and we find

$$\epsilon \sim \frac{u_\ell^3}{\ell}$$

It also follows that $\tau_\ell = \ell/u_\ell = \ell(\epsilon \ell)^{-1/3} = \epsilon^{-1/3} \ell^{2/3}$ so that the circulation time decreases with the length scale and becomes of the order of the viscous dissipation time $(\nu/\epsilon)^{1/2}$ precisely when $\ell \sim \ell_\epsilon$.

A similar relation between ϵ and the large scales can also be obtained with heuristic arguments: let e be the mean kinetic energy and ℓ_0 , a characteristic length for the large scales. Then u_0 given by $e = u_0^2/2$ is a characteristic velocity for the large scales, and $\tau_0 = \ell_0/u_0$ is the large-scale circulation time. In statistical equilibrium, the rate ϵ of kinetic energy dissipated per unit time and unit mass is expected to be of the order of e/τ_0 , hence

$$\epsilon \sim \frac{u_0^3}{\ell_0}$$

which is called the “energy dissipation law.”

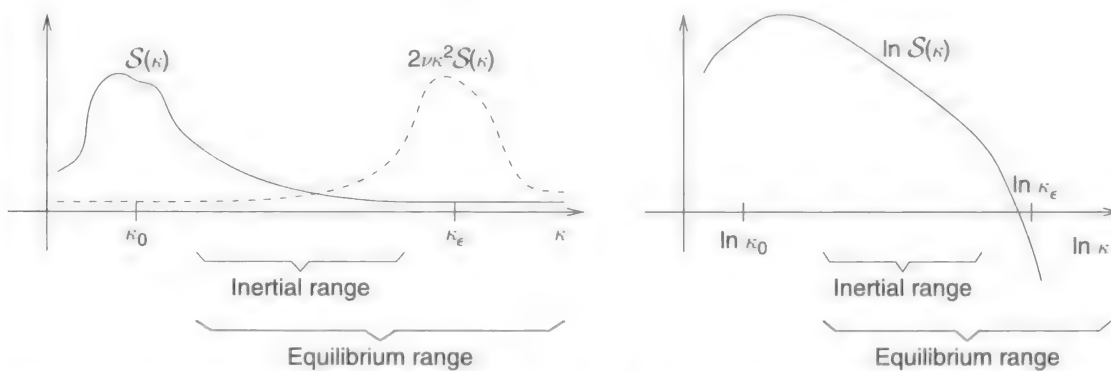


Figure 4 A typical distribution for the energy spectrum $S(\kappa)$ and the dissipation spectrum $2\nu\kappa^2 S(\kappa)$ in spectral space in nonlogarithmic and logarithmic scales. The energy is mostly concentrated on the large scales while the dissipation is concentrated near the dissipation scale. In the logarithmic scale, the four-fifths law for the energy spectrum stands out as a straight line with slope $-4/5$ over the inertial range.

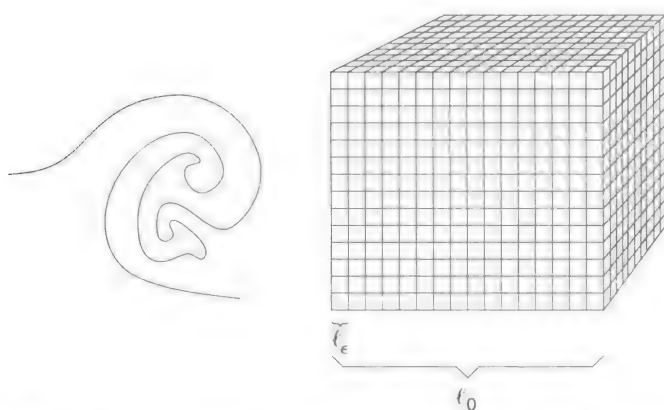


Figure 5 A schematic representation of a flow structure displaying a range of active scales and a three-dimensional grid with linear dimension l_0 and mesh length l_ϵ , sufficient to represent all the active scales in a turbulent flow. The number of degrees of freedom is the number of blocks: $(l_0/l_\epsilon)^3$.

From the energy dissipation law, several relations between characteristic quantities of turbulent flows can be obtained, such as $l_0/l_\epsilon \sim Re^{3/4}$, for $Re = l_0 u_0/\nu$.

Now, assuming the active scales in a turbulent flow exist down to the Kolmogorov scale l_ϵ , one needs a three-dimensional grid with mesh spacing l_ϵ to resolve all the scales, which means that the number N of degrees of freedom of the system is of the order of $N \sim (l_0/l_\epsilon)^3$ (see Figure 5). This number can be estimated in terms of the Reynolds number by $N \sim Re^{9/4}$. This relation is important in predicting the computational power needed to simulate all the active scales in turbulent flows.

Several such universal laws can be deduced and extended to other situations such as turbulent boundary layers, with the famous logarithmic law of the wall. They play a fundamental role in turbulence modeling and closure, for the calculation of the mean flow and other quantities.

Intermittency

The universality hypothesis based on a constant mean energy dissipation rate throughout the flow received some criticisms and was later modified by Kolmogorov in an attempt to account for observed large deviations on the mean rate of energy dissipation. Such phenomenon of intermittency is related to the vortex stretching and thinning mechanism, which leads to the formation of coherent structures of vortex filaments of high vorticity and low dissipation (Figure 6). These filaments have diameter as small as the Kolmogorov scale and longitudinal length extending from the Taylor scale up to the large scales and with a lifetime of the order of the large-scale circulation time.

It has been argued based on experimental evidence that intermittency leads to modified power laws

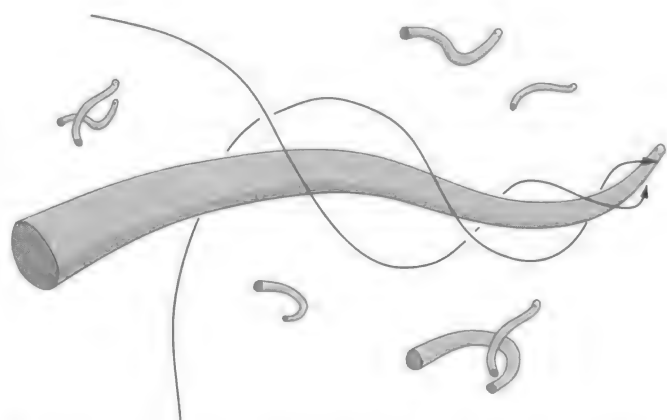


Figure 6 A portion of rotating fluid gets stretched and thinned as the flow speeds up, generating one of many coherent structures of high vorticity and low dissipation.

$S_p(\ell) \propto \ell^{\zeta(p)}$, $\zeta(p) < p/3$, for high-order ($p > 3$) structure functions. The issues of intermittency and coherent structures and whether and how they could affect the deductions of the universality theory such as the power laws for the structure functions are far from settled and are currently one of the major and most fascinating issues being addressed in turbulence theory. Several phenomenological theories attempt to adjust the universality theory to the existence of such coherent structures. Multifractal models, for instance, suppose that the eddies generated in the cascade process do not fill up the space and form multifractal structures. Field-theoretic renormalization group develops techniques based on quantum field renormalization theory. Intermediate asymptotics also exploits self-similar analysis and renormalization theory but with a somewhat different flavor. Detailed mathematical analysis of the vorticity equations is also playing a major role in the understanding of the dynamics of the vorticity field.

Mathematical Aspects of Turbulence Theory

From a mathematical perspective, it is fundamental to develop a rigorous background upon which to study the physical quantities of a turbulent flow. The first problem in the mathematical theory is related to the deterministic nature of chaotic systems assumed in dynamical system theory and believed to hold in turbulence. This has actually not been proved for the Navier–Stokes equations. It is in fact one of the most outstanding open problems in mathematics to determine whether given an initial condition for the velocity field there exists, in some sense, a unique solution of the Navier–Stokes equations starting with this initial condition and valid for all later times. It has been proved that a global solution (i.e., valid for all later

times) exists but which may not be unique, and it has been proved that unique solutions exist which may not be global (i.e., they are guaranteed to exist as unique solutions only for a finite time).

The difficulty here is the possible existence of singularities in the vorticity field (vorticity becoming infinite at some points in space and time). Depending on how large the singularity set is, uniqueness may fail in strictly mathematical terms. The existence of singularities may not be a purely mathematical curiosity, it may in fact be related with the intermittency phenomenon. Rigorous studies of the vorticity equation may continue to reveal more fundamental aspects on vortex dynamics and coherent structures.

The statistical theory has also been put into a firm foundation with the notion of statistical solution of the Navier–Stokes equations. It addresses the existence and regularity of the probability distribution assumed for turbulent flows and of the fundamental elements of the statistical theory such as correlation functions and spectra. Based on that, a number of relations between physical quantities of turbulent flows may be derived in a mathematically sound and definitive way. This does not replace other theories, it is mostly a mathematical framework upon which other techniques can be applied to yield rigorous results.

Despite the difficulties in the mathematical theory of the NSE some successes have been collected such as estimates for the number of degrees of freedom in terms of fractal dimensions of suitable sets associated with the solutions of the Navier–Stokes equations, and partial estimates of a number of relations derived in the statistical theory of fully developed turbulence.

See also: Bifurcations in Fluid Dynamics; Geophysical Dynamics; Incompressible Euler Equations: Mathematical Theory; Intermittency in Turbulence; Inviscid Flows; Lagrangian Dispersion (Passive Scalar); Stochastic Hydrodynamics; Variational Methods in

Turbulence; Viscous Incompressible Fluids: Mathematical Theory; Vortex Dynamics; Wavelets: Application to Turbulence.

Further Reading

- Adzhemyan LT, Antonov NV, and Vasiliev AN (1999) *The Field Theoretic Renormalization Group in Fully Developed Turbulence*. Amsterdam: Gordon and Breach.
- Barenblatt GI and Chorin AJ (1998) New perspectives in turbulence: scaling laws, asymptotics, and intermittency. *SIAM Review* 40(2): 265–291.
- Batchelor GK (1953) *The Theory of Homogeneous Turbulence*. Cambridge Monographs on Mechanics and Applied Mathematics. New York: Cambridge University Press.
- Chorin AJ (1994) Vorticity and Turbulence. *Applied Mathematical Sciences* vol. 103, New York: Springer.
- Constantin P (1994) Geometric statistics in turbulence. *SIAM Review* 36: 73–98.
- Foias C, Manley OP, Rosa R, and Temam R (2001) Navier–Stokes Equations and Turbulence. *Encyclopedia of Mathematics and its Applications*, vol. 83. Cambridge: Cambridge University Press.
- Friedlander S and Topper L (1961) *Turbulence. Classic Papers on Statistical Theory*. New York: Interscience Publisher.
- Frisch U (1995) *Turbulence. The Legacy of A. N. Kolmogorov*. Cambridge: Cambridge University Press.
- Hinze JO (1975) *Turbulence*. McGraw-Hill Series in Mechanical Engineering. New York: McGraw-Hill.
- Holmes P, Lumley JL, and Berkooz G (1996) *Turbulence, Coherent Structures, Dynamical Systems, and Symmetry*. Cambridge: Cambridge University Press.
- Lesieur M (1997) *Turbulence in Fluids. Fluid Mechanics and its Applications*, 3rd edn. vol. 40. Dordrecht: Kluwer Academic.
- Monin AS and Yaglom AM (1975) *Statistical Fluid Mechanics: Mechanics of Turbulence 2*. Cambridge: MIT Press.
- Sagaut P (2001) *Large Eddy Simulation for Incompressible Flows*. Berlin: Springer.
- Schlichting H and Gersten K (2000) *Boundary Layer Theory*, 8th edn. Berlin: Springer.
- Tennekes H and Lumley JL (1972) *A First Course in Turbulence*. Cambridge, MA: MIT Press.
- Vishik MI and Fursikov AV (1988) *Mathematical Problems of Statistical Hydrodynamics*. Dordrecht: Kluwer.
- Wilcox DC (2000) *Turbulence Modeling for CFD*, 2nd edn. Anaheim, CA: DCW Industries Inc.

Twistor Theory: Some Applications

L Mason, University of Oxford, Oxford, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Roger Penrose introduced twistor theory as a geometrical framework for basic physics in order to unify quantum theory and gravity. This program has had many successes along the way, but the long-term goals

of reformulating and superceding the established theories of basic physics are still a long way from being fulfilled. Nevertheless, the successes have had many important applications across mathematics and mathematical physics. This article will concentrate on three areas of application: integrable systems, geometry, and perturbative gauge theory (via twistor-string theory). It is intended to be self-contained as far as possible, but the reader may well find it easier to first read the article *Twistors*.

Twistor Theory

A basic motivation of twistor theory is to bring out the complex (holomorphic) geometry that underlies real spacetime. In general relativity, a spacetime is a 4-manifold with metric g of signature $(1, 3)$, and when it is flat, that is, $g = dt^2 - dx^2 - dy^2 - dz^2$, where (t, x, y, z) are coordinates on \mathbb{R}^4 , it is called Minkowski space. The first appearance of a complex structure arises from the fact that, at a given event, the celestial sphere of light rays (directions of zero length with respect to g) naturally has the structure of the Riemann sphere, \mathbb{CP}^1 , in such a way that Lorentz transformations (linear transformations of the tangent space preserving the metric) act on this sphere by Möbius transformations. These are the maximal group of complex analytic transformations of \mathbb{CP}^1 .

Twistor space extends this idea to the whole of Minkowski space. Denoted \mathbb{PT} , the twistor space for Minkowski space is complex projective 3-space, \mathbb{CP}^3 , the space of one-dimensional subspaces of \mathbb{C}^4 ; it is a three-dimensional complex manifold obtained by adding a "plane at infinity" to \mathbb{C}^3 . Explicitly, we can introduce homogeneous coordinates $Z^\alpha \in \mathbb{C}^4 - \{0\}$ with $\alpha = 0, 1, 2, 3$ but where $Z^\alpha \sim \lambda Z^\alpha$ for $\lambda \in \mathbb{C} - \{0\}$. Affine coordinates on a \mathbb{C}^3 chart $Z^3 \neq 0$ can be obtained by setting $(z_1, z_2, \lambda) = (Z^0/Z^3, Z^1/Z^3, Z^2/Z^3)$. Physically, points of twistor space correspond to spinning massless particles in Minkowski space. Mathematically, the correspondence can be understood as the Klein correspondence.

The Klein Correspondence

The correspondence between \mathbb{PT} and Minkowski space can be extended first to complexified Minkowski space so that the coordinates are allowed to take on values in \mathbb{C} , and then to its conformal compactification by including the "light cone at infinity." It then coincides with the classical complex Klein correspondence. The Klein correspondence is the one-to-one correspondence between lines in \mathbb{CP}^3 and points of a four complex-dimensional quadric, \mathbb{CM} , in \mathbb{CP}^5 . The 4-quadric \mathbb{CM} can be understood as conformally compactified complexified Minkowski space. Introducing affine coordinates (z_1, z_2, λ) on \mathbb{PT} and (t, x, y, z) on \mathbb{CM} , we find that a point (t, x, y, z) in \mathbb{CM} corresponds to a line in \mathbb{PT} according to

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} t - z & x + iy \\ x - iy & t + z \end{pmatrix} \begin{pmatrix} 1 \\ \lambda \end{pmatrix}$$

Alternatively, fixing (λ, z_1, z_2) in these equations gives a 2-plane in complex Minkowski space corresponding to all the lines in \mathbb{PT} through (λ, z_1, z_2) . Such 2-planes are called " α -planes."

They are totally null (i.e., the tangent vectors not only have zero length but are also mutually orthogonal) and also self-dual (under the differential geometer's notion of Hodge duality).

This complex correspondence can also be restricted to give correspondences for \mathbb{R}^4 with metrics of positive-definite, Euclidean, signature or ultrahyperbolic, $(2, 2)$, signature. A particular simplification in Euclidean signature is that the complex α -planes intersect the real slice in a point. The conformal compactification of Euclidean \mathbb{R}^4 is the 4-sphere S^4 given by adding a single point at infinity, and so we have a projection $p: \mathbb{PT} \rightarrow S^4$ whose fibers are holomorphically embedded \mathbb{CP}^1 s. These fibers can be characterized as the lines in \mathbb{PT} that are invariant under a quaternionic complex conjugation which is an antiholomorphic map $\hat{\cdot}: \mathbb{PT} \rightarrow \mathbb{PT}$ with no fixed points. (Here quaternionic means that on the nonprojective twistor space, $T = \mathbb{C}^4$, the conjugation has the property $\hat{Z} = -Z$ so that it defines a second complex structure anticommuting with the standard one; this is sufficient to express $T = \mathbb{Q}^2$, where \mathbb{Q} denotes the quaternions. The complex structures i, j , and k of the quaternions are given by identifying i with $\sqrt{-1}$ on \mathbb{C}^4 and j with $\hat{\cdot}$ and $k = ij$.)

The Penrose Transform

A basic task of twistor theory is to transform solutions to the field equations of mathematical physics into objects on twistor space. This works well for linear massless fields such as the Weyl neutrino equation, Maxwell's equations for electromagnetism and linearized gravity. In its general form, this transform has become known as the Penrose transform. Such fields correspond to freely prescribable holomorphic functions $f(\lambda, z_1, z_2)$ (or, more precisely, analytic cohomology classes) on regions of twistor space. The field can be obtained from this function by means of a contour integral. The simplest of these integral formulas is

$$\phi(x^a) = \oint f(\lambda, t - z + \lambda(x + iy), x - iy + \lambda(t + z)) d\lambda$$

and differentiation under the integral sign leads easily to the fact that ϕ satisfies the wave equation

$$\frac{\partial^2 \phi}{\partial t^2} - \frac{\partial^2 \phi}{\partial x^2} - \frac{\partial^2 \phi}{\partial y^2} - \frac{\partial^2 \phi}{\partial z^2} = 0$$

This formula was originally discovered by Bateman. Note that f must have singularities on twistor space to yield a nontrivial ϕ and even then, there are many choices of f that yield zero. For a solution ϕ defined

over a region U in spacetime, the function f is correctly understood as a representative of a Čech cohomology class defined on the region U' in twistor space swept out by the lines corresponding to points of U . Furthermore, the function f should be taken globally to be a function of homogeneity -2 , $f(\lambda Z^\alpha) = \lambda^{-2}f(Z^\alpha)$. This formula has generalizations to massless fields of all helicities in which a field of helicity s corresponds to a function (Čech cocycle) of homogeneity degree $2s - 2$.

The Penrose transform has found important applications in representation theory and integral geometry. For a review, the reader is referred to Baston and Eastwood (1989), the relevant survey articles in Bailey and Baston (1990), or Mason and Hughston (1990, chapter 1).

Twistor Theory and Nonlinear Equations

The Penrose transform for the Maxwell equations and linearized gravity turns out to be linearizations of correspondences for the nonlinear analogs of these equations: the Einstein vacuum equations and the Yang–Mills equations. However, the constructions only work when these fields are anti-self-dual. This is the condition that the curvature 2-forms satisfy $F^* = -iF$, where $*$ denotes the Hodge dual (which, up to certain factors of $\pm i$, has the effect of interchanging electric and magnetic fields); it is a nonlinear generalization of the right-handed circular polarization condition. Explicitly, in terms of spacetime indices $a, b, \dots = 0, 1, 2, 3$, $F_{ab}^* = (1/2)\epsilon_{abcd}F^{cd}$, where $\epsilon_{0123} = 1$ and $\epsilon_{abcd} = \epsilon_{[abcd]}$. In Minkowski signature, the i factor in the anti-self-duality condition implies that real fields cannot be anti-self-dual. Thus, these extensions are not sufficient to fulfill the ambitions of twistor theory to incorporate real classical nonlinear physics in Minkowski space. However, the factor of i is not present in Euclidean and ultrahyperbolic signature, so the anti-self-duality condition is consistent with real fields in these signatures and this is where the main applications of these constructions have been.

The Nonlinear Graviton Construction and Its Generalizations

The first nonlinear twistor construction was due to Penrose (1976), and was inspired by Newman's (1976) construction of "heavens" from the infinities of asymptotically flat spacetimes in general relativity.

The nonlinear graviton construction proceeds from the definition of twistors in flat spacetime as α -planes in complexified Minkowski space. It is

natural to ask which complexified metrics admit a full family of α -surfaces, that is, 2-surfaces that are totally null and self-dual. The answer is that a full family of α -surfaces exists iff the conformally invariant part of the curvature tensor, the Weyl tensor, is anti-self-dual. If this is the case, twistor space can be defined to be the (necessarily three-dimensional) space of such α -surfaces.

The remarkable fact is that the twistor space, together with its complex structure, is sufficient to determine the original spacetime. Twistor space is again a three-dimensional complex manifold, and contains holomorphically embedded rational curves, \mathbb{CP}^1 s, at least one for each point of the spacetime. However, holomorphic rigidity implies that the family of rational curves is precisely four-dimensional over the complex numbers. Furthermore, incidence of a pair of curves can be taken to imply that the corresponding points in spacetime lie on a null geodesic and this yields a conformal structure on spacetime. Further structures on twistor space can be imposed to give the complex spacetime a metric that is vacuum, perhaps with a cosmological constant. The correspondence is stable under small deformations and so the data defining the twistor space is effectively freely prescribable, see Penrose (1976).

In Euclidean signature, again the complex α -planes intersect the real spacetime in a point, so the twistor space again fibers over spacetime. The twistor fibration can be constructed as the projectivized bundle of self-dual spinors or more commonly as the unit sphere bundle in the space of self-dual 2-forms (Atiyah *et al.* 1978). In the latter formulation, the complex structure on the twistor space arises from the direct sum of the naturally defined complex structures on the horizontal and vertical tangent spaces to the bundle; that on the vertical subspace is the standard one on the sphere, and that on the horizontal subspace is a multiple of the self-dual 2-form at the given point of the fiber.

There are now large families of extensions, generalizations, and reductions of this construction. They are all based on the idea of realizing a space with a given complexified geometric structure as the parameter space of a family of holomorphically embedded submanifolds inside a twistor space. In general, the most useful of these constructions are those in which the "spacetime" is obtained as the space of rational curves in a twistor space. This is because the equations that are solved on the corresponding spacetime can be thought of as a completely integrable system in which the integrability condition for the generalized α -surfaces is interpreted as the consistency condition of a Lax

pair or more general linear system. For a more detailed discussion from this point of view, see Mason and Woodhouse (1996, chapter 13).

The Anti-Self-Dual Yang–Mills Equation and Its Twistor Correspondence

The anti-self-dual Yang–Mills equations extend Maxwell’s equations for electromagnetism in the right-circularly polarized case. They are a family of equations that depend on a choice of Lie group G , usually taken to be a group of complex matrices; Maxwell’s equations arise when $G = U(1)$.

Introduce coordinates $x^a, a = 0, 1, 2, 3$, on \mathbb{R}^4 with metric $ds^2 = dx^0 \cdot dx^3 - dx^1 \cdot dx^2$ (this is a metric of ultrahyperbolic signature – Euclidean signature can be obtained by choosing the coordinates to be complex, but with $(x^3, -x^2)$ the complex conjugates of (x^0, x^1)). The dependent variables are the components A_a of a connection $D_a = \partial_a - A_a$, where $\partial_a = \partial/\partial x^a$ and $A_a = A_a(x^b) \in \text{Lie } G$, the Lie algebra of G . This connection defines a method of differentiating vector-valued functions s in some representation of G . The freedom in changing bases for the vector bundle induce the gauge transformations $A_a \rightarrow g^{-1}A_ag - g^{-1}\partial_ag, g(x) \in G$ on A_a ; two connections that are related by a gauge transformation are deemed to be the same. The self-dual Yang–Mills equations are the condition

$$[D_0, D_2] = [D_1, D_3] = [D_0, D_3] - [D_1, D_2] = 0$$

They are the compatibility conditions

$$[D_0 + \lambda D_1, D_2 + \lambda D_3] = 0$$

for the linear system of equations

$$(D_0 - \lambda D_1)s = (D_2 - \lambda D_3)s = 0 \quad [1]$$

where $\lambda \in \mathbb{C}$ and s is an n -component column vector. These latter equations form a “Lax pair” for the system.

The Ward (1977) construction provides a one–one correspondence between gauge equivalence classes of solutions of the self-dual Yang–Mills equations and holomorphic vector bundles on regions in twistor space. The key point here is that eqn [1] defines parallel propagation along α -planes. To each point Z in twistor space, we can associate the vector space E_Z of solutions to eqn [1] along the corresponding α -plane. These vector spaces vary holomorphically with Z and that is what one means by a holomorphic vector bundle $E \rightarrow \text{PT}$. The remarkable fact is that the anti-self-dual Yang–Mills field can be reconstructed up to gauge from E , and, in effect, for local analytic solutions, E can be represented by freely prescribable “patching” data

consisting of local holomorphic matrix-valued functions on twistor space. To construct the solution on spacetime, one must first find a Birkhoff factorization of the patching data on each Riemann sphere in twistor space corresponding to points of the appropriate region in spacetime. On each Riemann sphere, the Birkhoff factorization starts with the given patching function with values in $GL(n, \mathbb{C})$ on the real axis in the complex plane, and expresses it as a product of functions with values in $GL(n, \mathbb{C})$ one of which extends over the upper-half plane, and the other over the lower-half complex plane. The anti-self-dual connection can be obtained by differentiating the resulting matrices. See Penrose (1984, 1986), Ward and Wells (1990), or Mason and Woodhouse (1996) for a full discussion, and Atiyah (1979) for the formulation appropriate to Euclidean signature.

Completely Integrable Systems

In effect, the twistor constructions amount to providing a geometric general local solution to the anti-self-duality equations; the twistor data is, for a local solution, freely prescribable. In this sense, they demonstrate complete integrability of the anti-self-duality equations. The reconstruction of a solution on spacetime from twistor data is not a quadrature – it involves, in the anti-self-dual Yang–Mills case, a Birkhoff factorization (also sometimes referred to as the solution to a Riemann–Hilbert problem), and in the case of the anti-self-dual Einstein equations, the construction of a family of rational curves inside a complex manifold. Nevertheless, such constructions are a familiar part of the apparatus of the theory of integrable systems.

In Ward (1985), this connection with integrable systems was developed further, and the anti-self-dual Yang–Mills equations were shown to yield many important integrable systems under symmetry reduction. Ward’s list has been extended and now includes many of the most famous examples of integrable systems such as the Painlevé equations, the Korteweg–de Vries (KdV) equation, the non-linear Schrödinger equation, the n -wave equations, and so on, see Mason and Woodhouse (1996) for a review. There are some notable omissions from the list such as the Kadomtsev–Petviashvili (KP) and Davey–Stewartson equations (at least if one restricts oneself to finite-dimensional gauge groups; reductions using infinite dimensional gauge groups have been obtained).

The list of integrable systems obtainable by symmetry reduction nevertheless remains impressive and provides a route to the classification of at least those integrable systems that can be obtained in this

way. Such systems can be classified by the choice of ingredients required in the symmetry reduction: the gauge group, the group of spacetime symmetries to be reduced by, the choice of Euclidean or ultra-hyperbolic signature, and the choice of certain constants of integration that arise in the reduction.

Another implication is that if an integrable system can be obtained from one of the self-duality equations by symmetry reduction, then it inherits a reduced twistor correspondence because the twistor correspondences share the symmetry groups of the spacetime field equations. These twistor correspondences can be seen to underlie much of the theory of these equations; for example, Backlund transformations of solutions correspond to elementary algebraic operations on the twistor data, similarly the Kac–Moody Lie algebras of hidden symmetries act locally on the twistor data by matrix multiplication of the appropriate loop algebras. Similarly, the inverse-scattering transform for the KdV and nonlinear Schrödinger equations can be seen to arise as particular presentations of the twistor construction.

By and large, although twistor methods have yielded new insight into the geometry and structure of systems in dimensions 1 and 2, they have not necessarily superseded pre-existing techniques for constructing solutions and analyzing the solution space. The systems for which twistor methods have been particularly effective for constructing solutions and characterizing their properties are in $2+1$ or higher dimension. Key examples here are of course the anti-self-dual Yang–Mills and Einstein equations themselves, and their single translation reductions. In the anti-self-dual Yang–Mills case, these reductions lead either to Ward’s or Manakov and Zakharov’s chiral model in Lorentzian signature, $2+1$, or the Bogomolny equations for monopoles, the reduction from Euclidean signature. In both cases, the twistor construction has played a major role in constructing and studying the solitonic solutions.

See Ward and Wells (1990), Mason and Woodhouse (1996), Ward’s article in Huggett *et al.* (1998) and the first few chapters of Mason *et al.* (1995), and Mason *et al.* (2001) for more examples of aspects of the theory of integrable systems arising from twistor correspondences.

Applications to Geometry

These applications are, to a large extent, higher-dimensional analogs of those discussed above; most of the problems in geometry to which twistor theory has been applied are those for which the underlying differential equations are integrable. These start

with the Euclidean signature versions of the original Ward construction for anti-self-dual Yang–Mills fields and Penrose’s nonlinear graviton construction for Ricci-flat anti-self-dual metrics but, as we will discuss, these constructions have a number of extensions and generalizations.

The first dramatic application of these constructions was the ADHM construction of Yang–Mills instantons. These are absolute minima of the Yang–Mills action, $S[A] = \int \text{tr}(F \wedge F^*)$ on the 4-sphere, S^4 , with its round metric. A simple argument shows that the action is bounded below by the second Chern class of the bundle and that this bound is achieved only for anti-self-dual fields. Thus, the problem was to characterize all the anti-self-dual Yang–Mills fields on S^4 . In this Euclidean context, twistor space, \mathbb{CP}^3 , fibers over S^4 and the corresponding Ward vector bundle is a bundle over all of \mathbb{CP}^3 . It turns out that all such bundles satisfying a certain stability condition had been constructed reasonably explicitly by algebraic geometers. Since the stability condition was implied by the context, this could be turned into an algebraic construction of the general instanton explicit enough to give some insight into both the local and global structure of the solution space. See Atiyah (1979) for a review.

Hitchin used the Euclidean version of the nonlinear graviton to develop the theory of gravitational instantons that are asymptotically locally Euclidean (i.e., asymptotically \mathbb{R}^4/Γ , where Γ is a finite subgroup of the rotation group). These were finally constructed by Kronheimer who again used twistor theory to identify the appropriate parameter space, see his article in Mason *et al.* (2001) and Dancer’s review of hyper-Kähler manifolds in LeBrun and Wang (1999).

Even in four dimensions, there are a number of variants of the nonlinear graviton construction. The basic twistor correspondence produces a twistor space that is a complex 3-manifold \mathcal{PT} for 4-manifolds with conformal structures whose Weyl tensor is anti-self-dual. There are four natural specializations that have attracted study: (1) the Ricci-flat case, (2) the Einstein case (with nonzero cosmological constant), (3) the scalar-flat Kähler case, and (4) the hypercomplex case.

The twistor space in the Ricci-flat case admits the additional structure of a fibration over \mathbb{CP}^1 together with a holomorphic Poisson structure on the fibers with values in the pullback of the 1-forms on \mathbb{CP}^1 (alternatively, the bundle of holomorphic 3-forms should be the pullback of the square of the bundle of holomorphic 1-forms on \mathbb{CP}^1). The Einstein case with nonzero cosmological constant is a variant of this in which the twistor space admits a

nondegenerate holomorphic contact structure, that is, a distribution of 2-plane elements, which are only integrable when the cosmological constant vanishes. It also admits a Kähler form when the scalar curvature is positive (in the negative case the corresponding Kähler form is indefinite). For the case of Kähler metrics with vanishing scalar curvature, the twistor space admits a holomorphic volume form with a double pole. The Ricci-flat case is equivalent to the case of hyper-Kähler metrics, those that are Kähler with respect to three different complex structures I, J , and K satisfying the standard quaternionic relations $IJ = K$, etc. A hypercomplex structure is obtained when one only has the three integrable complex structures satisfying the quaternion relations. Such manifolds admit an underlying conformal structure that is anti-self-dual, and the corresponding twistor space admits a fibration to \mathbb{CP}^1 .

These constructions have all played a significant role in the general analysis of these geometric structures, and the construction of examples. A striking example of an application of the nonlinear graviton construction to general properties is due to Donaldson and Friedman who show that if two 4-manifolds admit anti-self-dual conformal structures, then their direct sum does also.

In higher dimensions, most generalizations rely on quaternionic geometry and its reductions. The Euclidean signature formulation of the nonlinear graviton construction has natural extensions to quaternionic manifolds in $4k$ dimensions. These are manifolds with metric whose holonomies are contained in $\mathrm{Sp}(k) \times \mathrm{Sp}(1)$. The latter $\mathrm{Sp}(1) = \mathrm{SU}(2)$ factor leads to an associated S^2 bundle whose total space is the twistor space \mathcal{PT} and it naturally has the structure of a $(2k+1)$ -dimensional complex manifold.

For a series of review articles, the reader is referred to Bailey and Baston (1990, chapters 3 and 4) and also LeBrun and Wang (1999, chapters 2, 5, 6, 10, and 14) which, despite being a book on the distinct subject of Einstein manifolds, is strongly influenced by twistor theory. Other applications along these lines are summarized in Mason *et al.* (2001, chapter 1).

There are a number of applications that go beyond complete integrability. A striking application is the twistor framework of Merkulov for studying arbitrary geometric structures. This has led to a classification of all possible irreducible holonomies of torsion-free affine connections, see Merkulov's article in Huggett *et al.* (1998). Another important area is in the field of conformal invariants in which the local twistor connection plays a prominent role.

This is a connection that is naturally defined on any conformal manifold being the spinor representation of the Cartan conformal connection. An impressive application here is the construction of conformally invariant differential operators and other conformal invariants. See the article by Baston and Eastwood in Bailey and Baston (1990).

Beyond Classical Integrability: Twistor-String Theory

Until Witten (2004), there was little indication that twistor theory would have much useful to say about Yang–Mills or gravitational fields that are not anti-self-dual. Furthermore, it was problematic to incorporate quantum field theory into twistor ideas. However, twistor-string theory has transformed the situation and has furthermore had impressive applications to the field of perturbative gauge theory.

The story starts with a formulation by Nair of the remarkable Park–Taylor formulas for the so-called maximal helicity violating (MHV) amplitudes in gauge theory. These are scattering amplitudes at tree level in which helicity conservation is maximally violated; using crossing symmetry to take all the particles to be outgoing, these are amplitudes in which $n-2$ of the particles have helicity -1 and two have helicity $+1$. These amplitudes can be expressed simply as follows. Let the n particles have color t_i in the Lie algebra of the gauge group and null momenta p_i with spinor decompositions $p_i^a = \tilde{\pi}_i^A \pi_i^{A'}$, $i=1, \dots, n$ where the $\pi_i^{A'}$ are self-dual spinors and $\tilde{\pi}_i^A$ are anti-self-dual spinors using the index notation of Spinors and Spin Coefficients, and Twistors. Let $i=r$ and $i=s$ be the two gluons of helicity $+1$. Then the coefficient of the colour term $\mathrm{tr}(t_1 t_2 \cdots t_n)$ is

$$\delta^4 \left(\sum_{i=1}^n p_i^a \right) \frac{\pi_r \cdot \pi_s}{\prod_{i=1}^n \pi_i \cdot \pi_{i+1}}$$

where $\pi_i \cdot \pi_j = \pi_i^{A'} \pi_{jA'}$ denotes the standard skew-symmetric inner product on chiral spinors and $\pi_{n+1} = \pi_1$. A striking feature is that, except for the delta function, it is holomorphic in the π_i s except at the simple poles $\pi_i \cdot \pi_{i+1} = 0$. Nair interprets these poles as those associated to fermion correlators in a current algebra on a \mathbb{CP}^1 parametrized by π . Using a supersymmetric formulation adapted to $N=4$ super Yang–Mills, he formulated the amplitude as arising from an integral over lines in supertwistor space $\mathbb{CP}^{3|4}$.

Witten extends these ideas to give, at least conjecturally, a complete theory. He proposes that full perturbative $N=4$ super Yang–Mills theory on spacetime is equivalent to a string theory, a topological

B model, on a supersymmetric version of twistor space, $\text{PT}_s = \mathbb{CP}^{3|4}$. This is the space obtained by taking $\mathbb{C}^{4|4}$ with bosonic coordinates $Z^\alpha, \alpha = 0, \dots, 3$ and fermionic coordinates $\eta^i, i = 1, \dots, 4$ moduli the equivalence relation $(Z^\alpha, \eta^i) \sim \lambda(Z^\alpha, \eta^i)$ where $\lambda \in \mathbb{C}, \lambda \neq 0$.

The number 4 here plays two crucial but different roles. It is the maximum number of supersymmetries that Yang–Mills can have; it has the effect of incorporating both the positive and negative helicity parts of the gauge field in the same supermultiplet. It is also the only value of N for which $\mathbb{CP}^{3|N}$ is a Calabi–Yau manifold and this is a necessary condition for the topological twisted B model to be anomaly-free. The Calabi–Yau condition is the condition that the manifold admit a global holomorphic volume form which here is

$$\Omega_s = \varepsilon_{\alpha\beta\gamma\delta} Z^\alpha dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta \wedge d\eta^1 \wedge d\eta^2 \wedge d\eta^3 \wedge d\eta^4$$

This is invariant under $(Z^\alpha, \eta^i) \rightarrow (\lambda Z^\alpha, \lambda \eta^i)$ because $d(\lambda \eta^i) = \lambda^{-1} d\eta^i, \lambda \in \mathbb{C}$ follows from the Berezinian rule of integration $\int \theta d\theta = 1$ for anticommuting variables.

Open-string topological twisted B models are known to correspond to holomorphic Chern–Simons theories on their target space. A holomorphic Chern–Simons theory is a theory whose basic variable is a $\bar{\partial}$ -operator $\bar{\partial}_A = \bar{\partial} + A$ on a complex vector bundle $E \rightarrow \text{PT}^{3|4}$, where A is a Lie algebra valued $(0, 1)$ -form on the target space and whose action is

$$S[A] = \int \frac{1}{2} \left(A \bar{\partial} A + \frac{1}{3} A^3 \right) \wedge \Omega_s$$

The field equations are $\bar{\partial}_A^2 = 0$. The classical solutions therefore consist of holomorphic vector bundles on the target space, here $\mathbb{CP}^{3|4}$. The twistor-space representation of the fields are obtained by expanding A in the anticommuting variables η^i to obtain

$$A = a + \eta^i b_i + \eta^i \eta^j c_{ij} + \eta^i \eta^j \eta^k d_{ijk} + \eta^1 \eta^2 \eta^3 \eta^4 g$$

and a has homogeneity zero, but because the homogeneity of η^i is of degree 1, b_i has homogeneity degree -1 , and so on down to homogeneity degree -4 for g . Via the Ward construction, the a component corresponds to an anti-self-dual Yang–Mills field on spacetime. The other components of A can be seen to correspond to spacetime fields with helicities $-1/2$ to $+1$ that are background coupled to the anti-self-dual Yang–Mills field.

As it stands, although this holomorphic Chern–Simons theory gives the correct field content of $N=4$ super Yang–Mills, the couplings are only those of an anti-self-dual sector and more couplings are needed to obtain full $N=4$ super Yang–Mills. The remarkable fact is that these can be naturally introduced by coupling in certain D1 instantons. The D1 instantons are algebraic curves C in twistor space and the coupling is via a pair of spinor fields α and β on C with values in E and E^* , respectively with action

$$S[\alpha, \beta, A] = \int_C \beta \bar{\partial}_A \alpha$$

This leads to explicit expressions for Yang–Mills scattering amplitudes in terms of integrals of fermion correlators over the moduli spaces of such algebraic curves in supertwistor space. In principle, the integral is over all algebraic curves. However, algebraic curves have two topological invariants, their degree denoted d and genus g . An argument based on a classical scaling symmetry gives that integration over just those of curves of degree d gives the subset of processes for which

$$d = q - 1 + l$$

where q is the number of outgoing particles of helicity $+1$ in the process and l is the number of loops. It is also the case that $g \leq l$.

An elegant formula for the amplitudes is that for the on-shell generating functional for tree-level scattering amplitudes $\mathcal{M}^d[A]$, where A is the on-shell twistor field, being the above-mentioned $(0, 1)$ -form. The generating functional for processes with $q = d + 1$ external fields of helicity $+1$ is then

$$\mathcal{M}^d[A] = \int_{C \in \mathcal{M}^d} \det(\bar{\partial} + A)|_C d\mu$$

where $d\mu$ is a natural measure on the moduli space \mathcal{M}^d of connected rational (genus 0) curves in $\mathbb{CP}^{3|4}$ of degree d . This approach has been successfully exploited to obtain implicit algebraic formulas for all tree-level scattering amplitudes.

In an alternative version, the curves of degree d can be taken to be maximally disconnected, being the union of d lines. However, in this approach, we need to also incorporate Chern–Simons propagators which, for tree diagrams, join the lines into a tree. This gives a very flexible calculus for perturbative gauge theory in which scattering processes are obtained by gluing together MHV diagrams. It has been argued that the two formulations are equivalent. On the one hand, the Chern–Simons propagator has a simple pole when the lines meet and the

contour integral over the moduli space can be performed using residues in such a way as to eliminate the Chern–Simons propagators leaving an integral over d intersecting lines. On the other hand, the measure on the space of connected curves has a simple pole where the curve acquires double points and again the contour integral can be performed in such a way as to yield the same integral over d intersecting lines.

It should be mentioned that Berkovits has given an alternative version of twistor-string theory which is a heterotic open-string theory with target supertwistor space in which the strings are taken to have boundary on the real slice \mathbb{RP}^3 in \mathbb{CP}^3 (this is appropriate to a spacetime with split signature) and the D1-instanton expansions are replaced by expansions in the fundamental modes of the string (this is not a topological theory). This gives rise to the same formulas for scattering amplitudes as Witten’s original model.

There have been many applications now of these ideas, perhaps the most striking being the recursion relations of Britto, Cachazo, Feng, and Witten which give, at tree level, on-shell recurrence relations for Yang–Mills scattering amplitudes that suggests a hitherto unsuspected underlying structure for Yang–Mills theory.

Despite all these successes, twistor-string theory is not thought by string theorists to be a good vehicle for basic physics. The most serious problem is that the closed-string sector gives rise to conformal supergravity which is an unphysical theory. This is particularly pernicious from the point of view of analyzing loop diagrams as from the point of view of string theory, loop diagrams will carry supergravity modes. From this point of view, twistor-string theory is another duality, like AdS–CFT etc., that gives insight into some standard physics but is fundamentally limited.

From the point of view of a twistor theorist, however, twistor-string theory has overcome major obstacles to the twistor programme. Hodges has used the BCFW recursion relations to provide all twistor diagrams for gauge theory. In Mason (2005) it is shown how to derive the main generating function formulas from Yang–Mills and conformal gravity spacetime action principles via a twistor space actions for these theories. These twistor actions can in the first instance be expressed purely bosonically and distinctly and the twistor-string generating function formulas are obtained by expanding and re-summing the classical limit of the path integral in a parameter that expands about the anti-self-dual sector. This allows one to decouple the Yang–Mills and conformal gravity modes, and indeed to work purely bosonically – one is not tied to super Yang–Mills. Although there is much work

to be done to extend these ideas to provide a consistent approach to the main equations of basic physics, obstacles that seemed insurmountable a few years ago have been overcome.

See also: Chern–Simons Models: Rigorous Results; Einstein Equations: Exact Solutions; General Relativity: Overview; Instantons: Topological Aspects; Integrable Systems and the Inverse Scattering Method; Riemann–Hilbert Methods in Integrable Systems; Spinors and Spin Coefficients; Twistors; Classical Groups and Homogeneous Spaces; Quantum Mechanics: Foundations; Several Complex Variables: Compact Manifolds; Several Complex Variables: Basic Geometric Theory.

Further Reading

- Atiyah MF (1979) *Geometry of Yang–Mills Fields: Lezioni Fermiane*. Pisa: Accademia Nazionale dei Lincei Scuola Normale Superiore.
- Atiyah MF, Hitchin NJ, and Singer IM (1978) Self-duality in four-dimensional Riemannian geometry. *Proceedings of the Royal Society A* 362: 425.
- Bailey TN and Baston R (eds.) (1990) *Twistors in Mathematics and Physics*, LMS Lecture Notes Series, vol. 156. Cambridge: Cambridge University Press.
- Baston RJ and Eastwood MG (1989) *The Penrose Transform: Its Interaction with Representation Theory*. Oxford: Oxford University Press.
- Cachazo F, and Svrcek P (2005) Lectures on twistor strings and perturbative Yang–Mills theory, arXiv:hep-th/0504194.
- Hitchin N (1987) *Monopoles, Minimal Surfaces and Algebraic Curves*, *Seminaire de Mathématiques supérieures*, vol. 105. NATO Advanced Study Institute. Les Presses de l’Université De Montreal.
- Huggett S, Mason LJ, Tod KP, Tsou TS, and Woodhouse NMJ (eds.) (1998) *The Geometric Universe*. Oxford: Oxford University Press.
- LeBrun C and Wang M (eds.) (1999) *Essays on Einstein manifolds*, *Surveys in Differential Geometry*, vol. VI. Boston, MA: International Press.
- Mason LJ (2005) Twistor actions for non-self-dual fields, a derivation of twistor-string theory, hep-th/0507269.
- Mason LJ and Hughston LP (eds.) (1990) *Further Advances in Twistor Theory, Volume I: The Penrose Transform and Its Applications*. Pitman Research Notes in Maths, vol. 231. Harlow: Longman.
- Mason LJ, Hughston LP, and Kobak PZ (1995) *Further Advances in Twistor Theory, Volume II: Integrable Systems, Conformal Geometry and Gravitation*. Pitman Research Notes in Maths, vol. 232. Harlow: Longman.
- Mason LJ, Hughston LP, Kobak PZ, and Pulverer K (eds.) (2001) *Further Advances in Twistor Theory, Volume III: Curved Twistor Spaces*. Pitman Research Notes in Maths, vol. 424. Boca Raton, FL: Chapman and Hall/CRC Press.
- Mason LJ and Woodhouse NMJ (1996) *Twistor Theory, Self-Duality and Twistor Theory*. Oxford: Oxford University Press.
- Newman ET (1976) Heaven and its properties. *General Relativity and Gravitation* 7(1): 107–111.
- Penrose R (1976) Nonlinear gravitons and curved twistor theory. *General Relativity and Gravitation* 7: 31–52.

- Penrose R (1984, 1986) *Spinors and Space-Time*, vols. I and II. Cambridge: Cambridge University Press.
- Ward RS (1977) On self-dual gauge fields. *Physics Letters A* 61: 81–82.
- Ward RS (1985) Integrable and solvable systems and relations among them. *Philosophical Transactions of the Royal Society A* 315: 451–457.

- Ward RS and Wells RO (1990) *Twistor Geometry and Field Theory*. Cambridge: Cambridge University Press.
- Witten E (2004) Perturbative gauge theory as a string theory in twistor space. *Communications in Mathematical Physics* 252: 189 (arXiv:hep-th/0312171).

Twistors

K P Tod, University of Oxford, Oxford, UK

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Twistor theory initially arose from two principal motivations: a desire for a conformally invariant calculus for spacetime geometry and fields on spacetime, and a desire to unify and account for the various occurrences of complex numbers and holomorphic functions in mathematical physics, especially in general relativity (Penrose and MacCallum 1973). The theory leads to a nonlocal relation between spacetime and twistor space, whereby a point in one is an extended object in the other. Part of the present-day motivation of the subject is that this nonlocal relation will be a fruitful way to approach the quantization of spacetime. A comparison is often invoked with Hamiltonian mechanics, which is a formal rephrasing of classical mechanics that nonetheless provides a bridge from that theory to quantum mechanics. The hope is that the twistor theory has the right character to provide a bridge from general relativity to quantum theory, specifically to quantum gravity.

The principal successes of twistor theory in mathematical physics can be characterized as the linear Penrose transform, which provides a solution of the zero-rest-mass free-field equations in Minkowski space in terms of sheaf cohomology in twistor space, and the nonlinear Penrose transform, which provides solutions of certain nonlinear field equations in terms of holomorphic geometry. These are treated below, together with other applications of twistor theory, following a brief introduction to twistor geometry.

Very recently, there has been a resurgence of interest in twistor theory following Witten's introduction of twistor string theory (Witten 2003) as a string theory in twistor space. This is not treated here, but this article does provide the necessary background.

Twistor Geometry

General references for this section are the books by Penrose and Rindler (1986) and Hugget and Tod (1994). It will be convenient to use Penrose's abstract index convention (Penrose and Rindler 1984, 1986), which is also used in *Spinors and Spin Coefficients*. This can be used wherever vector or tensor indices occur. Suppose that V is a (real or complex) finite-dimensional vector space with dual V' . Elements of V are written v^a, u^b, w^c, \dots , where an index a, b, c, \dots is regarded not as an integer in the range 1 to $\dim V$ but simply as an abstract label indicating that the object to which it is attached is a vector. Elements of V' are similarly written u_a, v_b, w_c, \dots and elements of the tensor algebra as $t^{a\dots b}_{c\dots d}$ according to valence, and so on. The usual operations of tensor algebra are written in the way that component calculations would suggest, but without necessitating a choice of basis. The jump to tensor fields on a manifold M is immediate. A metric is a particular field g_{ab} and determines a Levi-Civita connection ∇_a which defines maps $\nabla_a: v^b \rightarrow \nabla_a v^b$ and similar for other valences. The virtue of the formalism is that, while remaining invariant, it can harness the strength and flexibility of calculations in components.

With this understanding, twistors may first be defined as the fundamental representation of $SU(2, 2)$, so that they are elements Z^α of a four-dimensional complex vector space T . T carries a Hermitian form Σ of signature $(++--)$ which is made explicit below and which provides an isomorphism from the complex conjugate of T to its dual. This isomorphism is used to eliminate all appearances of complex-conjugate twistors from the formalism and is therefore regarded as an antilinear map to the dual.

$SU(2, 2)$ is the double cover of $O(2, 4)$, the rotation group of $E_{2,4}$, the six-dimensional space with flat metric $\eta_{2,4}$ of signature $(+---+)$, which in turn is the double cover of $C(1, 3)$, the conformal group of Minkowski space M . This last group homomorphism may be made explicit as follows (suspending the abstract-index convention for the duration of this

aside): introduce pseudo-Cartesian coordinates $x^a = (x^0, x^1, x^2, x^3)$ on M and $y^\alpha = (y^0, y^1, y^2, y^3, y^4, y^5)$ on $E_{2,4}$. The corresponding metrics are

$$\begin{aligned} \eta_{1,3} &= \eta_{ab} dx^a dx^b \\ &= (dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2 \end{aligned} \quad [1]$$

$$\begin{aligned} \eta_{2,4} &= \eta_{\alpha\beta} dy^\alpha dy^\beta \\ &= (dy^0)^2 - (dy^1)^2 - (dy^2)^2 - (dy^3)^2 \\ &\quad + (dy^4)^2 - (dy^5)^2 \end{aligned} \quad [2]$$

We map M into $E_{2,4}$ by

$$\phi(x^a) = (x^0, x^1, x^2, x^3, (1 - \eta)/2, (1 + \eta)/2) \quad [3]$$

where $\eta = \eta_{ab} x^a x^b$ with η_{ab} as in [1], and it can be checked that $\phi(M)$ is the intersection of the null cone N of the origin in $E_{2,4}$ with the plane P defined by $y^4 + y^5 = 1$. P is in fact a null hyperplane in $E_{2,4}$ and any point of N not on the null hyperplane defined by

$$y^4 + y^5 = 0 \quad [4]$$

can be mapped along the generators of N to a unique point of P (recall that any point on a cone lies on a line through the vertex: these lines are the generators). Thus, the image of M under ϕ gives a point on every generator of N except those satisfying [4]. It can also be seen from [2] that the intrinsic metric in $E_{2,4}$ on the intersection of N and P is just $\eta_{1,3}$.

Now let PN be the projective null cone, or, equivalently, the space of generators of N . This is a compact manifold with topology $S^1 \times S^3$, as one can see by intersecting N with the sphere

$$(y^0)^2 + (y^1)^2 + (y^2)^2 + (y^3)^2 + (y^4)^2 + (y^5)^2 = 2$$

Each generator meets this sphere twice at, say, y^α and $-y^\alpha$, and PN is the quotient by this identification of the two surfaces

$$(y^0)^2 + (y^4)^2 = 1 = (y^1)^2 + (y^2)^2 + (y^3)^2 + (y^5)^2$$

which define the intersection. The metric $\eta_{2,4}$ defines a degenerate metric on N , which, however, is nondegenerate on any smooth cross section of N which meets each generator once. Furthermore, the map along the generators between any two such cross sections is conformal. Thus, there is a conformal metric on PN and it is conformal to $\eta_{1,3}$. We call PN compactified Minkowski space M_c as it is compact and has the same conformal metric as Minkowski space. It can be thought of as M compactified by the addition of some points, namely

the points of PN corresponding to the generators satisfying [4]. To interpret these, we consider the points satisfying the similar equation $y^4 - y^5 = 0$. By inspection of ϕ , [3], we see that these points correspond to the light cone of the origin in M . Thus, M_c is obtained from M by adding a single light cone, the light cone at infinity known as \mathcal{I} and read as “scri,” short for “script-I.”

Now the rotation group $O(2, 4)$ of $E_{2,4}$ maps N to itself preserving the metric and consequently maps PN to itself, preserving the conformal metric. Thus, $O(2, 4)$ defines conformal transformations of M_c and a count of dimension shows that it is locally isomorphic to the conformal group $C(1, 3)$. The map is two-to-one with $\pm I$ in $O(2, 4)$ mapping to I in $C(1, 3)$. The fact that $SU(2, 2)$ is four-to-one homomorphic to $C(1, 3)$ follows from calculations below. It is because of this homomorphism of $SU(2, 2)$ and $C(1, 3)$ that the geometry and analysis of twistors (i.e., twistor theory) provides a formalism adapted to conformally invariant or conformally covariant notions in M or M_c .

A twistor may be expressed in terms of two-component spinors of $SL(2, \mathbb{C})$, the double cover of the Lorentz group, as follows:

$$Z^\alpha = (\omega^A, \pi_{A'}) \quad [5]$$

where again indices are abstract, so that

$$\mathbb{T} = \mathbb{S} \oplus \bar{\mathbb{S}}'$$

in terms of the spin space \mathbb{S} and complex-conjugate dual spin space $\bar{\mathbb{S}}'$ of M . Now we can write the action of infinitesimal elements of $C(1, 3)$ explicitly as

$$\begin{aligned} \hat{\omega}^A &= \phi^A_B \omega^B - i T^{AA'} \pi_{A'} + \Lambda \omega^A \\ \hat{\pi}_{A'} &= \bar{\phi}^{B'}_{A'} \pi_{B'} + i B_{AA'} \omega^A + \Lambda \pi_{A'} \end{aligned} \quad [6]$$

where $T^{AA'}$ (a real vector) defines an infinitesimal translation, $B_{AA'}$ (another real vector) defines an infinitesimal special conformal transformation, Λ (a real constant) defines a dilatation and the (real) bivector $M_{ab} = \phi_{AB} \epsilon_{A'B'} + \bar{\phi}_{A'B'} \epsilon_{AB}$ defines an infinitesimal rotation. This gives a total of 15 parameters for the transformation, which is the correct dimension for $C(1, 3)$.

The Hermitian form $\Sigma(,)$ can be written as

$$\Sigma(Z, Z) = Z^\alpha \bar{Z}_\alpha = \omega^A \bar{\pi}_A + \bar{\omega}^{A'} \pi_{A'} \quad [7]$$

when it can be checked that the transformations [6] leave it invariant (and that its signature is $(+ + - -)$; this establishes that $SU(2, 2)$ is locally isomorphic to $C(1, 3)$). Equation [7] will be referred to as the norm of a twistor.

From [6], a twistor $Z^\alpha = (\omega^A, \pi_{A'})$ gives rise, under translation by a variable $x^{AA'}$, to a spinor field Ω^A given by

$$\Omega^A = \omega^A - ix^{AA'}\pi_{A'} \quad [8]$$

Differentiating [8] and symmetrizing, we see that Ω^A satisfies the differential equation

$$\nabla_{A'(A}\Omega_{B)} = 0 \quad [9]$$

which is known as the twistor equation. In fact, the general solution of [9] takes the form of [8] for constant spinors ω^A and $\pi_{A'}$. Furthermore, the conformal group can be shown directly to act on solutions of [9], so that twistor theory can begin with the study of [9] and its solutions. In this approach, a twistor is precisely a solution of [9].

Given a spinor field Ω^A of the form of [8], we may seek the points of M where it vanishes. In general, there are none, but if we consider complexified Minkowski space CM , then Ω^A vanishes on a two-dimensional complex plane with the property that every tangent vector is of the form $\lambda^A\pi_{A'}$ for varying λ^A and fixed $\pi_{A'}$. The 2-plane is flat and totally null, in that the (analytically extended) Minkowski metric vanishes identically on it, and it has a self-dual (SD) tangent bivector determined by $\pi_{A'}$. Such a 2-plane is known as an α -plane (reserving the term β -plane for a totally null 2-plane with anti-self-dual (ASD) tangent bivector). At a given point p in CM , there is an α -plane for each choice of $\pi_{A'}$ up to scale (in other words, for each element of the projective (primed) spin space at p) which is a copy of the complex projective line, CP^1 .

The α -plane is determined by the twistor up to scale (in that a constant complex multiple of the field Ω^A determines the same α -plane). Thus, we consider the projective twistor space PT which, since T is C^4 , is a copy of complex projective 3-space, CP^3 . This is now the space of α -planes, but is also compact. We define complexified, compactified Minkowski space CM_c as the space of all (complex projective) lines in PT ; then it is easy to see that this includes CM as an open dense subset. PT is the space of α -planes in CM_c and two lines meet in PT iff the corresponding points in CM_c lie on an α -plane, or, equivalently, iff they are null separated. Thus, the conformal structure in CM_c is determined by incidence of lines in PT .

To find M and M_c in this picture, we seek α -planes containing real points. If Ω^A from [8] vanishes at a real $x^{AA'}$, then the contraction $\omega^A\bar{\pi}_A$ must be purely imaginary, so that, by [7], the norm of the twistor is zero. Conversely, one calculates that Ω^A can indeed vanish at real points if the norm is zero, and that it will then in fact vanish along a null geodesic with

tangent vector (proportional to) $\bar{\pi}^A\pi^{A'}$. Twistors with norm zero are called null and the (five-dimensional, real) submanifold of them in PT is PN . This is a compactification of the space of (unscaled) null geodesics in M by the inclusion of the 2-sphere of null geodesics in M_c which lie on the light cone at \mathcal{I} . For use in the next section, we note the definition of PT^+ and PT^- as the projective twistors with positive and negative norm, respectively.

To summarize, we have found M and M_c : (complex projective) lines in PT define points of CM_c ; lines in PN define points of M_c with one such, call it I , picked out as the vertex of the null cone \mathcal{I} ; lines in PN which meet I correspond to points of \mathcal{I} ; lines in PN which do not meet I correspond to points in M . As for CM_c , the conformal structure of M and M_c is determined by incidence in PN . We may now note the nonlocal correspondence mentioned in the introduction: points in CM_c are lines in PT and points in PT are α -planes in CM_c .

It will be convenient to refer to the line in PT associated with a point x in CM_c as L_x . With this notation, it is possible to characterize the forward or future tube in terms of twistor space: a point x of CM is in the forward tube iff its imaginary part is timelike and past-pointing, and this is equivalent to L_x lying in PT^- .

The starting point for Riemannian twistor theory is the fact that CP^3 is a fibration with fiber CP^1 over S^4 , where the fiber above a point p can be interpreted as the almost-complex structures at p (since this is the same as the projective primed spin space at p). In the picture developed above, this means that there is an S^4 's worth of lines filling out CP^3 , no two of which intersect (so that there are no null vectors and the metric is definite). The complexification of S^4 with its conformal structure is again CM_c .

If a twistor has nonzero norm, say $Z^\alpha\bar{Z}_\alpha = s \neq 0$, then it can be interpreted as a massless particle with spin s : the momentum is $p_a = \bar{\pi}_A\pi_{A'}$ and the angular momentum bivector is $M^{ab} = i\omega^{(A}\bar{\pi}^{B)}\epsilon^{A'B'} - i\bar{\omega}^{(A'}\pi^{B')}\epsilon^{AB}$. The angular momentum transforms appropriately under translation by virtue of [6] and the (Pauli–Lubanski) spin vector is sp_a , as it should be for a massless spinning particle.

The Linear Penrose Transform: Zero-Rest-Mass Free Fields

A zero-rest-mass free field of spin s is a symmetric spinor field $\phi_{AB\dots C}$ with $2s$ indices which satisfies the field equation

$$\nabla^{A'A}\phi_{AB\dots C} = 0 \quad [10]$$

The Weyl neutrino equation, source-free Maxwell equation, and linearized Einstein vacuum equation are examples of zero-rest-mass free-field equations, with spins $1/2$, 1 , and 2 , respectively, so that these are equations of physical interest. Conventionally, one takes the $s=0$ case to be the wave equation, and the complex-conjugate fields $\psi_{A'B'...C'}$ to have the same spin but opposite helicity.

The conformal group acts on solutions of [10], so that the equations are conformally invariant. The equations can be solved by contour integral expressions involving homogeneous functions of a twistor variable. To be explicit, we define an operation ρ_x of restriction to the line L_x for a function of a twistor variable by the following:

$$\rho_x f(Z^\alpha) = f(ix^{AA'} \pi_{A'}, \pi_{A'}) \quad [11]$$

Now suppose that $f(Z^\alpha)$ is holomorphic and homogeneous of degree $-2s-2$ in the twistor variable for positive integer $2s$, but otherwise arbitrary, and consider the integral

$$\begin{aligned} &\psi_{A'B'...C'}(x) \\ &= \int \pi_{A'} \pi_{B'} \dots \pi_{C'} \rho_x f(Z^\alpha) \epsilon^{E'F'} \pi_{E'} d\pi_{F'} \quad [12] \end{aligned}$$

where there are $2s$ indices on ψ and the integration is around a contour in the line L_x in PT. The choice of homogeneity ensures that the integral is well defined but, to obtain a nonzero answer, $\rho_x f$ must have some singularities as a function of $\pi_{A'}$ on L_x . The answer then automatically gives a helicity- $(-s)$ solution of [10], as may be checked by differentiation under the integral sign.

For a helicity- s solution, we take an arbitrary function $f(Z^\alpha)$, holomorphic and of homogeneity $(2s-2)$, and consider the integral

$$\begin{aligned} &\phi_{AB...C}(x) \\ &= \int \rho_x \left(\frac{\partial}{\partial \omega^A} \frac{\partial}{\partial \omega^B} \dots \frac{\partial}{\partial \omega^C} f(Z^\alpha) \right) \epsilon^{E'F'} \pi_{E'} d\pi_{F'} \quad [13] \end{aligned}$$

where there are $2s$ indices on ϕ and the integration is again around a contour in the line L_x . As before, one needs singularities to make the contour integral nonzero, but again the result satisfies [10].

The correct framework in which to understand these integrals is sheaf cohomology theory. For [12], the functions with singularities are actually elements of $H^1(\hat{\mathcal{U}}, \mathcal{O}(-2s-2))$, the first cohomology group of a region $\hat{\mathcal{U}}$ in PT with coefficients in the sheaf of germs of holomorphic functions of homogeneity $-2s-2$, while the fields are elements of $H^0(\mathcal{U}, \mathcal{Z}_s)$, the zeroth cohomology group of the corresponding region \mathcal{U} of M with coefficients in

helicity- s zero-rest-mass fields (thus, $\hat{\mathcal{U}}$ must contain the neighborhood of lines L_x for points x in \mathcal{U}). Similarly, [13] is interpreted cohomologically in terms of potentials modulo a gauge. With appropriate conditions on $\hat{\mathcal{U}}$ and \mathcal{U} (for brevity, \mathcal{U} is said to be elementary), these groups can be shown to be isomorphic and this isomorphism is known as the Penrose transform (Ward and Wells (1991)). A particular instance of an elementary \mathcal{U} is the forward tube, when $\hat{\mathcal{U}}$ is \mathbb{PT}^- . Since the definition of positive frequency is holomorphicity on the forward tube, this observation geometrizes the notion of positive frequency in terms of twistor space.

For free fields with mass, there are generalizations of [12] and [13] to solve the Dirac equation for different spins. However, the integrands now involve functions of more than one twistor variable, subject to an equation. This equation is a counterpart of the Klein-Gordon equation and breaks the conformal invariance (as it must, since mass does). It can be imposed by a projection which can in turn be written as a contour integral over arbitrary holomorphic functions. It has been argued that the appropriate description of leptons and hadrons in twistor theory is with functions of two and three twistor variables, respectively. Such a function has two or three integer quantum numbers determined by the homogeneities in different variables, and this leads to a twistor particle classification scheme (see, e.g., Hughston and Sheppard (1980) and Sparling (1981)), similar in many respects to, but not identical with, the standard classifications.

Given that free fields, massive or massless, are determined from arbitrary twistor functions through contour integrals, one may translate the Feynman diagrams of a quantum field theory into contour integrals over twistor functions. In the massless case, the contours are compact, so that the integrals are finite without need for renormalization. The massive case is more complicated but essentially parallel. This is twistor diagram theory and there is a substantial literature on it (see, e.g., the article by Hodges in the volume edited by Huggett *et al.* (1998)). There is currently no new physical theory, distinct from a known quantum field theory, to generate the relevant diagrams.

The Nonlinear Penrose Transform: Curved Twistor Spaces

The electromagnetic field, in Minkowski space say, can be regarded as a spinor field subject to field equations, in which case these equations can be

solved via the Penrose transform by contour integrals. Alternatively, it can be seen as the curvature of a connection on a $U(1)$ bundle over M , which is a more active role for the field in curving a bundle. For SD or ASD electromagnetic fields, there are analogous active twistor constructions. From an ASD electromagnetic field, one may define a connection on the primed spin space of CM which is flat on α -planes: if the tangents to the α -plane are of the form $\lambda^A \pi^{A'}$ for varying λ^A and with $\pi^{A'}$ fixed up to scale, then consider the propagation of $\pi_{A'}$ around the α -plane given by

$$\pi^{A'}(\nabla_{A'A} - iA_{A'A})\pi_{B'} = 0 \quad [14]$$

where $A_{A'A}$ is a potential for the electromagnetic field. This connection is flat provided

$$\pi^{A'}\pi^{B'}\nabla_{AA'}A_{B'}^A = 0 \quad [15]$$

and if this is to hold for all $\pi_{A'}$ then $\nabla_{A(A'}A_{B')}^A$ vanishes and the electromagnetic field, defined as usual as the exterior derivative of the potential, is necessarily ASD. Now the space of α -planes in CM is projective twistor space PT , so we define a holomorphic C^* bundle \mathcal{T} over PT by taking the fiber above an α -plane to be choices of $\pi_{A'}$ scaled as in [14]. If we restrict attention to the α -planes through a given point p of CM, then by comparing the scalings at p we can trivialize the bundle; thus, \mathcal{T} is trivial on lines in PT . There is a converse to this construction and we have: *there is a one-to-one correspondence between holomorphic C^* bundles on a region \hat{U} in PT which are trivial on lines and ASD electromagnetic fields on the corresponding region U of CM (for elementary U).*

This construction can be extended to solve the ASD Yang–Mills equations with holomorphic vector bundles replacing holomorphic line bundles: *with \hat{U} and elementary U as above, there is a natural one-to-one correspondence between ASD $GL(n, C)$ gauge fields on U and holomorphic rank- n vector bundles \mathcal{E} over \hat{U} which are trivial on L_x for every x in U .*

ASD Yang–Mills fields cannot be real on M , but using Riemannian twistor theory, one can impose appropriate reality and globality conditions to ensure that these ASD Yang–Mills fields are both real and globally defined on S^4 . These are then instantons. The Atiyah–Drinfeld–Hitchin–Manin (ADHM) construction of instantons (Atiyah *et al.* 1978) proceeds via construction of the corresponding holomorphic vector bundles over twistor space.

The construction of ASD Yang–Mills fields is also the starting point for the twistor theory of integrable systems (Mason and Woodhouse 1996), following the observation that many of the known

completely integrable partial differential equations (PDEs) (including the sine-Gordon, Korteweg–de Vries (KdV) and nonlinear Schrödinger equations) are reductions of the ASD Yang–Mills equations. Solutions of these other integrable systems can be given in terms of a geometrical construction, usually of some structure in holomorphic geometry.

The other major active twistor construction, which historically preceded the Yang–Mills one, is Penrose’s nonlinear graviton (Penrose 1976), which solves the ASD Einstein vacuum equations. For this, one starts from a complex, four-dimensional manifold \mathcal{M} with holomorphic metric, vanishing Ricci curvature and ASD Weyl tensor. These conditions on the curvature are necessary and sufficient to allow the existence of α -surfaces, which generalize α -planes. They are two-dimensional totally null (complex) surfaces with SD tangent bivector, one for each choice of (null) SD bivector, or, equivalently, for each choice of primed spinor, at each point.

The space of α -surfaces is a three-dimensional complex manifold, the curved twistor space PT . This is curved inasmuch as it is not now (part of) CP^3 , but it still contains complex projective lines: given a point p in \mathcal{M} there is an α -surface through p for every primed spinor at p up to scale; these α -surfaces make up a projective line L_p in PT . The conditions on the curvature are equivalent to the statement that the Levi-Civita connection is flat on primed spinors, so that there exist constant primed spinors in \mathcal{M} , and the tangent bivector to an α -surface can be taken to be constant, without loss of generality. The map associating a constant primed spinor with each α -surface defines a projection π from PT to CP^1 , so that PT is a fibration over CP^1 . The lines L_p define a four-parameter family of sections of this fibration.

To define the metric of \mathcal{M} from PT , one needs the notion of normal bundle: the normal bundle of a submanifold Y in a manifold X is $N = TX|_Y / TY$ in terms of the tangent bundles TX and TY . The normal bundle \mathcal{N}_p of a particular section L_p is the same in PT as it was in PT , namely $H \oplus H$, where H is the hyperplane-section line bundle over CP^1 (Ward and Wells 1991). A section S_V of \mathcal{N}_p corresponds to a vector V in $T_p\mathcal{M}$ (think of it as an infinitesimally neighboring point in \mathcal{M}) and V is defined to be null iff S_V has a zero. Because of the nature of \mathcal{N} , this defines a quadratic conformal metric, which, furthermore, agrees with the conformal metric on \mathcal{M} and generalizes the definition of conformal metric for CM_c in terms of incidence in PT . To define the actual metric, as opposed to just the conformal metric, one has a covariant-constant

choice of $\epsilon^{A'B'}$ in \mathcal{M} which defines an ϵ on the base of the fibration, and a Poisson structure on the fibers μ of the projection. The definition of μ is more intricate, but the two structures enable the metric of \mathcal{M} to be recovered from PT . Penrose (1976) and Huggett and Tod (1994) provide more details.

Now the metric and curvature properties of \mathcal{M} are coded into holomorphic properties of PT together with ϵ and μ . These properties characterize \mathcal{M} : *subject to topological conditions on \mathcal{M} , there is a one-to-one correspondence between holomorphic solutions \mathcal{M} of the Einstein vacuum equations with ASD Weyl tensor and three-dimensional complex manifolds PT fibered over CP^1 , with a four-parameter of sections, each with normal bundle $H \oplus H$, and the forms ϵ and μ as above.*

In fact, one only needs to assume the existence of one section with the correct normal bundle and the full four-parameter family will automatically exist, at least near to the initial one. Penrose (1976) showed how curved twistor spaces with the necessary structures could be obtained by deforming the neighborhood of a line in the “flat” twistor space PT . The Kodaira–Spencer theory of complex deformations ensures that the necessary lines continue to exist under this deformation.

The original nonlinear graviton construction has been extended in various ways including the following: to allow the possibility of a cosmological constant (Ward and Wells 1991); to produce real, Riemannian solutions (Hitchin 1995); to solve other but related field equations (e.g., those for hypercomplex metrics, scalar-flat Kahler metrics or Einstein–Weyl structures).

The search for a twistor construction of the SD Einstein equations (distinct from a construction in terms of dual twistors, which is, of course, provided by deforming dual twistor space) is an active area of research. This and other applications of twistor theory, including a quasilocal definition of mass in general relativity, the classification of affine holonomies and the construction of four-dimensional conformal field theories, may be found in the literature cited in the “Further reading” section.

See also: Classical Groups and Homogeneous Spaces; Clifford Algebras and Their Representations; Integrable

Systems: Overview; Quantum Field Theory: A Brief Introduction; Quantum Mechanics: Foundations; Relativistic Wave Equations Including Higher Spin Fields; Riemann–Hilbert Problem; Spinors and Spin Coefficients; Twistor Theory: Some Applications.

Further Reading

- Atiyah MF, Hitchin NJ, Drinfeld VG, and Manin YuI (1978) Construction of instantons. *Physics Letters A* 65: 185–187.
- Bailey TN and Baston RJ (eds.) (1990) *Twistors in Mathematics and Physics*, LMS Lecture Note Series, vol. 156. Cambridge: Cambridge University Press.
- Frauenfelder J and Penrose R (2001) Twistors and general relativity. In: Engquist B and Schmid W (eds.) *Mathematics Unlimited – 2001 and Beyond*. Berlin: Springer.
- Hitchin NJ (1995) Twistor spaces, Einstein metrics and isomonodromic deformations. *Journal of Differential Geometry*. 42: 30–112.
- Hitchin NJ, Segal GB, and Ward RS (1999) *Integrable Systems. Twistors, Loop Groups, and Riemann Surfaces*, Oxford Graduate Texts in Mathematics, vol. 4. Oxford: Oxford University Press.
- Huggett SA and Tod KP (1994) *An Introduction to Twistor Theory*, LMS Student Text, vol. 4. Cambridge: Cambridge University Press.
- Huggett SA, Mason LJ, Tod KP, Tsou TS, and Woodhouse NMJ (eds.) (1998) *The Geometric Universe – Science, Geometry and the Work of Roger Penrose*. Oxford: Oxford University Press.
- Hughston LP and Sheppard M (1980) On the magnetic moments of hadrons. *Reports on Mathematical Physics* 18: 53–66.
- Mason LJ, Hughston LP, Kobak PZ, and Pulverer K (eds.) (1995, 1998, 2001) *Further Advances in Twistor Theory*, vols. 1–3. Boca Raton, FL: Pitman Advanced Publishing Programme and Chapman and Hall, CRC.
- Mason LJ and Woodhouse NMJ (1996) *Integrability, Self-Duality, and Twistor Theory*, LMS Monographs Series. Oxford: Clarendon.
- Penrose R (1976) Nonlinear gravitons and curved twistor theory. *General Relativity and Gravitation* 7: 31–52.
- Penrose R (1999) The central programme of twistor theory. *Chaos Solitons Fractals* 10: 581–611.
- Penrose R and MacCallum MAH (1973) Twistor theory: an approach to the quantisation of fields and space–time. *Physics Reports* 6C: 241–316.
- Penrose R and Rindler W (1984, 1986) *Spinors and Space–Time*, vols. 1 and 2. Cambridge: Cambridge University Press.
- Sparling GAJ (1981) Theory of massive particles. I. Algebraic structure. *Philosophical Transactions of the Royal Society of London A* 301: 27–74.
- Ward RS and Wells RO (1991) *Twistor Geometry and Field Theory*. Cambridge: Cambridge University Press.
- Witten E (2003) Perturbative gauge theory as a string theory in twistor space, hep-th/0312171.

Two-Dimensional Conformal Field Theory and Vertex Operator Algebras

M R Gaberdiel, ETH Zürich, Zürich, Switzerland

© 2006 Elsevier Ltd. All rights reserved.

Introduction

For the last twenty years or so, two-dimensional (2D) conformal field theories have played an important role in different areas of modern theoretical physics. One of the main applications of conformal field theory has been in string theory (see Compactification of Superstring Theory), where the excitations of the string are described, from the point of view of the world sheet, by a 2D conformal field theory. Conformal field theories have also been studied in the context of statistical physics, since the critical points of second-order phase transition are typically described by a conformal field theory. Finally, conformal field theories are interesting solvable toy models of genuinely interacting quantum field theories.

From an abstract point of view, conformal field theories are (Euclidean) quantum field theories that are characterized by the property that their symmetry group contains, in addition to the Euclidean symmetries, local conformal transformations, that is, transformations that preserve angles but not necessarily lengths. The local conformal symmetry is of special importance in two dimensions since the corresponding symmetry algebra is infinite dimensional in this case. As a consequence, 2D conformal field theories have an infinite number of conserved quantities, and are essentially solvable by symmetry considerations alone. The mathematical formulation of these symmetries has led to the concept of a vertex operator algebra, which has become a new branch of mathematics in its own right. In particular, it has played a major role in the explanation of “monstrous moonshine” for which Richard Borcherds received the Fields medal in 1998.

In the following, we want to explain the main features of conformal field theory using an algebraic approach that will naturally lead to the concept of a vertex operator algebra. There are other approaches to the subject, most notably the formulation, pioneered by Segal, of conformal field theory as a functor from the category of Riemann surfaces to the category of vector spaces. Due to limitations of space, however, we will not be able to discuss any of these other approaches here.

The Conformal Symmetry Group

The conformal symmetry group of the n -dimensional Euclidean space \mathbb{R}^n consists of the (locally defined) transformations that preserve angles but not necessarily lengths. The transformations that preserve angles as well as lengths are the well-known translations and rotations. The conformal group contains (in any dimension) in addition the dilatations or scale transformations

$$x^\mu \mapsto \tilde{x}^\mu = \lambda x^\mu \quad [1]$$

where $\lambda \in \mathbb{R}$ and $x^\mu \in \mathbb{R}^n$, as well as the so-called special conformal transformations,

$$x^\mu \mapsto \tilde{x}^\mu = \frac{x^\mu + x^2 a^\mu}{1 + 2(x \cdot a) + x^2 a^2} \quad [2]$$

where $a^\mu \in \mathbb{R}^n$ and $x^2 = x^\mu x_\mu$. (Note that this last transformation is only defined for $x^\mu \neq -a^\mu/a^2$.)

If the dimension n of the space \mathbb{R}^n is larger than 2, one can show that the full conformal group is generated by these transformations. For $n=2$, however, the group of (locally defined) conformal transformations is much larger. To see this, it is convenient to introduce complex coordinates for $(x, y) \in \mathbb{R}^2$ by defining $z = x + iy$ and $\bar{z} = x - iy$. Then any (locally) analytic function $f(z)$ defines a conformal transformation by $z \mapsto f(z)$, since analytic maps preserve angles. (Incidentally, the same also applies to $z \mapsto \bar{f}(\bar{z})$, but this would reverse the orientation.) Clearly, the group of such transformations is infinite dimensional; this is a special feature of two dimensions.

In this complex notation, the transformations that are generated by translations, rotations, dilatations, and special conformal transformations simply generate the Möbius group of automorphisms of the Riemann sphere

$$z \mapsto f(z) = \frac{az + b}{cz + d} \quad [3]$$

where a, b, c, d are complex constants with $ad - bc \neq 0$; since rescaling a, b, c, d by a common complex number does not modify [3], the Möbius group is isomorphic to $\text{SL}(2, \mathbb{C})/\mathbb{Z}_2$. In addition to these transformations (that are globally defined on the Riemann sphere), we have an infinite set of infinitesimal transformations generated by $L_n: z \mapsto z + \epsilon z^{n+1}$ for all $n \in \mathbb{Z}$. The generators $L_{\pm 1}$ and L_0 generate the

subgroup of Möbius transformations, and their commutation relations are simply

$$[L_m, L_n] = (m - n)L_{m+n} \tag{4}$$

In fact, [4] describes also the commutation relations of all generators L_n with $n \in \mathbb{Z}$: this is the Lie algebra of (locally defined) 2D conformal transformations – it is called the Witt algebra.

The General Structure of Conformal Field Theory

A 2D conformal field theory is determined (like any other field theory) by its space of states and the collection of its correlation functions (vacuum expectation values). The space of states is a vector space \mathcal{H} (which, in many interesting examples, is a Hilbert space), and the correlation functions are defined for collections of vectors in some dense subspace of \mathcal{H} . These correlation functions are defined on a 2D (Euclidean) space. We shall mainly be interested in the case where the underlying 2D space is a closed compact surface; the other important case concerning surfaces with boundaries (whose analysis was pioneered by Cardy) will be reviewed elsewhere (see the article Boundary Conformal Field Theory). The closed surfaces are classified (topologically) by their genus g , which counts the number of handles; the simplest such surface which we shall mainly consider is the sphere with $g=0$, the surface with $g=1$ is the torus, etc.

One of the special features of conformal field theory is the fact that the theory is naturally defined on a Riemann surface (or complex curve), that is, on a surface that possesses suitable complex coordinates. In the case of the sphere, the complex coordinates can be taken to be those of the complex plane that cover the sphere except for the point at infinity; complex coordinates around infinity are defined by means of the coordinate function $\gamma(z)=1/z$ that maps a neighborhood of infinity to a neighborhood of zero. With this choice of complex coordinates, the sphere is usually referred to as the Riemann sphere, and this choice of complex coordinates is, up to Möbius transformations, unique. The correlation functions of a conformal field theory that is defined on the sphere are thus of the form

$$\langle 0|V(\psi_1; z_1, \bar{z}_1) \cdots V(\psi_n; z_n, \bar{z}_n)|0\rangle \tag{5}$$

where $V(\psi, z, \bar{z})$ is the field that is associated to the state ψ , and z_i and \bar{z}_i are complex conjugates of one another. Here $|0\rangle$ denotes the $SL(2, \mathbb{C})/Z_2$ -invariant vacuum. The usual locality assumption of a 2D

(bosonic) Euclidean quantum field theory implies that these correlation functions are independent of the order in which the fields appear in [5].

It is conventional to think of $z=0$ as describing “past infinity,” and $z=\infty$ as “future infinity”; this defines a time direction in the Euclidean field theory and thus a quantization scheme (radial quantization). Furthermore, we identify the space of states with the space of “incoming” states; thus, the state ψ is simply

$$\psi = V(\psi; 0, 0)|0\rangle \tag{6}$$

We can think of z_i and \bar{z}_i in [5] as independent variables, that is, we may relax the constraint that \bar{z}_i is the complex conjugate of z_i . Then we have two commuting actions of the conformal group on these correlations functions: the infinitesimal action on the z_i variables is described (as before) by the L_n generators, while the generators for the action on the \bar{z}_i variables are \bar{L}_n . In a conformal field theory, the space of states \mathcal{H} thus carries two commuting actions of the Witt algebra. The generator $L_0 + \bar{L}_0$ can be identified with the time-translation operator, and thus describes the energy operator. The space of states of the physical theory should have a bounded energy spectrum, and it is thus natural to assume that the spectrum of both L_0 and \bar{L}_0 is bounded from below; representations with this property are usually called positive-energy representations. It is relatively easy to see that the Witt algebra does not have any unitary positive-energy representations except for the trivial representation. However, as is common in many instances in quantum theory, it possesses many interesting projective representations. These projective representations are conventional representations of the central extension of the Witt algebra

$$[L_m, L_n] = (m - n)L_{m+n} + \frac{c}{12} m(m^2 - 1)\delta_{m,-n} \tag{7}$$

which is the famous Virasoro algebra. Here c is a central element that commutes with all L_m ; it is called the central charge (or conformal anomaly).

Given the actions of the two Virasoro algebras (that are generated by L_n and \bar{L}_n), one can decompose the space of states \mathcal{H} into irreducible representations as

$$\mathcal{H} = \bigoplus_{ij} M_{ij} \mathcal{H}_i \otimes \bar{\mathcal{H}}_j \tag{8}$$

where $\mathcal{H}_i(\bar{\mathcal{H}}_j)$ denotes the irreducible representations of the algebra of $L_n(\bar{L}_n)$, and $M_{ij} \in \mathbb{N}_0$ describe the multiplicities with which these combinations of representations occur. (We are assuming here that the space of states is completely reducible with

respect to the action of the two Virasoro algebras; examples where this is not the case are the so-called logarithmic conformal field theories.) The positive-energy representations of the Virasoro algebra are characterized by the value of the central charge, as well as the lowest eigenvalue of L_0 ; the state ψ whose L_0 eigenvalue is smallest is called the highest-weight state, and its eigenvalue $L_0\psi = h\psi$ is the conformal weight. The conformal weight determines the conformal transformation properties of ψ : under the conformal transformation $z \mapsto f(z)$, $\bar{z} \mapsto \bar{f}(\bar{z})$, we have

$$V(\psi; z, \bar{z}) \mapsto (f'(z))^h (\bar{f}'(\bar{z}))^{\bar{h}} V(\psi; f(z), \bar{f}(\bar{z})) \quad [9]$$

where $L_0\psi = h\psi$ and $\bar{L}_0\psi = \bar{h}\psi$. The corresponding field $V(\psi; z, \bar{z})$ is then called a primary field; if [9] only holds for the Möbius transformations [3], the field is called quasiprimary.

Since L_m with $m > 0$ lowers the conformal weight of a state (see [7]), the highest-weight state ψ is necessarily annihilated by all L_m (and \bar{L}_m) with $m > 0$. However, in general the L_m (and \bar{L}_m) with $m < 0$ do not annihilate ψ ; they generate the descendants of ψ that lie in the same representation. Their conformal transformation property is more complicated, but can be deduced from that of the primary state [9], as well as the commutation relations of the Virasoro algebra.

The Möbius symmetry (whose generators annihilate the vacuum) determines the 1-, 2- and 3-point functions of quasiprimary fields up to numerical constants: the 1-point function vanishes, unless $h = \bar{h} = 0$, in which case $\langle 0|V(\psi; z, \bar{z})|0\rangle = C$, independent of z and \bar{z} . The 2-point function of ψ_1 and ψ_2 vanishes unless $h_1 = h_2$ and $\bar{h}_1 = \bar{h}_2$; if the conformal weights agree, it takes the form

$$\begin{aligned} \langle 0|V(\psi_1; z_1, \bar{z}_1)V(\psi_2; z_2, \bar{z}_2)|0\rangle \\ = C(z_1 - z_2)^{-2h}(\bar{z}_1 - \bar{z}_2)^{-2\bar{h}} \end{aligned} \quad [10]$$

Finally, the structure of the 3-point function of three quasiprimary fields ψ_1, ψ_2 , and ψ_3 is

$$\begin{aligned} \langle 0|V(\psi_1; z_1, \bar{z}_1)V(\psi_2; z_2, \bar{z}_2)V(\psi_3; z_3, \bar{z}_3)|0\rangle \\ = C \prod_{i < j} (z_i - z_j)^{(h_k - h_i - h_j)} (\bar{z}_i - \bar{z}_j)^{(\bar{h}_k - \bar{h}_i - \bar{h}_j)} \end{aligned} \quad [11]$$

where for each pair $i < j$, k labels the third field, that is, $k \neq i$ and $k \neq j$. The Möbius symmetry also restricts the higher correlation function of quasiprimary fields: the 4-point function is determined up to an (undetermined) function of the Möbius invariant cross-ratio, and similar statements also

hold for n -point functions with $n \geq 5$. The full Virasoro symmetry must then be used to restrict these functions further; however, since the generators L_n with $n \leq -2$ do not annihilate the vacuum $|0\rangle$, the Virasoro symmetry leads to Ward identities that cannot be easily evaluated in general. (In typical examples, the Ward identities give rise to differential equations that must be obeyed by the correlation functions.)

Chiral Fields and Vertex Operator Algebras

The decomposition [8] usually contains a special class of states that transform as the vacuum state with respect to \bar{L}_m ; these states are the so-called chiral states. (Similarly, the states that transform as the vacuum state with respect to L_m are the antichiral states.) Given the transformation properties described above, it is not difficult to see that the corresponding chiral fields $V(\psi; z, \bar{z})$ only depend on z in any correlation function, that is $V(\psi; z, \bar{z}) \equiv V(\psi; z)$. (Similarly, the antichiral fields only depend on \bar{z} .) The chiral fields always contain the field corresponding to the state $L_{-2}|0\rangle$, that describes a specific component of the stress-energy tensor.

In conformal field theory, the product of two fields can be expressed again in terms of the fields of the theory. The conformal symmetry restricts the structure of this operator product expansion:

$$\begin{aligned} V(\psi_1; z_1, \bar{z}_1)V(\psi_2; z_2, \bar{z}_2) \\ = \sum_i (z_1 - z_2)^{\Delta_i} (\bar{z}_1 - \bar{z}_2)^{\bar{\Delta}_i} \\ \sum_{r,s \geq 0} V(\phi_{r,s}^i; z_2, \bar{z}_2) (z_1 - z_2)^r (\bar{z}_1 - \bar{z}_2)^s \end{aligned} \quad [12]$$

where Δ_i and $\bar{\Delta}_i$ are real numbers, and $r, s \in \mathbb{N}_0$. (Here i labels the conformal representations that appear in the operator-product expansion, while r and s label the different descendants.) The actual form of this expansion (in particular, representations that appear) can be read off from the correlation functions of the theory since the identity [12] has to hold in all correlation functions.

Given that the chiral fields only depend on z in all correlation functions, it is then clear that the operator-product expansion of two chiral fields again only contains chiral fields. Thus, the subspace of chiral fields closes under the operator-product expansion, and therefore defines a consistent (sub)theory by itself. This subtheory is sometimes referred to as a meromorphic conformal field theory (Goddard 1989). (Obviously, the same also applies to the subtheory of antichiral fields.) The operator-product

expansion defines a product on the space of meromorphic fields. This product involves the complex parameters z_i in a nontrivial way, and therefore does not directly define an algebra structure; it is, however, very similar to an algebra, and is therefore usually called a vertex operator algebra in the mathematical literature. The formal definition involves formal power series calculus and is quite complicated; details can be found in (Frenkel–Lepowski–Meurman 1988).

By virtue of its definition as an identity that holds in arbitrary correlation functions, the operator-product expansion is associative, that is,

$$(V(\psi_1; z_1, \bar{z}_1)V(\psi_2; z_2, \bar{z}_2))V(\psi_3; z_3, \bar{z}_3) \\ = V(\psi_1; z_1, \bar{z}_1)(V(\psi_2; z_2, \bar{z}_2)V(\psi_3; z_3, \bar{z}_3)) \quad [13]$$

where the brackets indicate which operator-product expansion is evaluated first. If we consider the case where both ψ_1 and ψ_2 are meromorphic fields, then the associativity of the operator-product expansion implies that the states in \mathcal{H} form a representation of the vertex operator algebra. The same also holds for the vertex operator algebra associated to the anti-chiral fields. Thus the meromorphic fields encode in a sense the symmetries of the underlying theory: this symmetry always contains the conformal symmetry (since $L_{-2}|0\rangle$ is always a chiral field, and $\bar{L}_{-2}|0\rangle$ always an antichiral field). In general, however, the symmetry may be larger. In order to take full advantage of this symmetry, it is then useful to decompose the full space of states \mathcal{H} not just with respect to the two Virasoro algebras, but rather with respect to the two vertex operator algebras; the structure is again the same as in [8], where, however, each \mathcal{H}_i and $\bar{\mathcal{H}}_i$ is now an irreducible representation of the chiral and antichiral vertex operator algebra, respectively.

Rational Theories and Zhu's Algebra

Of particular interest are the rational conformal field theories that are characterized by the property that the corresponding vertex operator algebras only possess finitely many irreducible representations. (The name “rational” stems from the fact that the conformal weights and the central charge of these theories are rational numbers.) The simplest example of such rational theories are the so-called minimal models, for which the vertex operator algebra describes just the conformal symmetry: these models exist for a certain discrete set of central charges $c < 1$ and were first studied by Belavin, Polyakov, and Zamolodchikov in 1984. (Their paper is contained in the reprint volume of Goddard and Olive (1988).) It was this seminal

paper that started many of the modern developments in conformal field theory. Another important class of examples are the Wess–Zumino–Witten (WZW) models that describe the world-sheet theory of strings moving on a compact Lie group. The relevant vertex operator algebra is then generated by the loop group symmetries. There is some evidence that all rational conformal field theories can be obtained from the WZW models by means of two standard constructions, namely by considering cosets and taking orbifolds; thus rational conformal field theory seems to have something of the flavor of (reductive) Lie theory.

Rational theories may be characterized in terms of Zhu's algebra that can be defined as follows. The chiral fields $V(\psi, z)$ that only depend on z must by themselves define local operators; they can therefore be expanded in a Laurent expansion as

$$V(\psi, z) = \sum_{n \in \mathbb{Z}} V_n(\psi) z^{-n-h} \quad [14]$$

where h is the conformal weight of the state ψ . For example, for the case of the holomorphic component of the stress–energy tensor one finds

$$T(z) = \sum_{n \in \mathbb{Z}} L_n z^{-n-2} \quad [15]$$

where the L_n are the Virasoro generators. By the state/field correspondence [6], it then follows that

$$V_n(\psi)|0\rangle = 0 \quad \text{for } n > -h \quad [16]$$

and that

$$V_{-h}(\psi)|0\rangle = \psi \quad [17]$$

(For an example of the above component of the stress–energy tensor, [16] implies that $L_{-1}|0\rangle = L_0|0\rangle = L_n|0\rangle = 0$ for $n \geq 0$ – thus the vacuum is in particular $\text{SL}(2, \mathbb{C})/\mathbb{Z}_2$ invariant. Furthermore, [17] shows that $L_{-2}|0\rangle$ is the state corresponding to this component of the stress–energy tensor.) We denote by \mathcal{H}_0 the space of states that can be generated by the action of the modes $V_n(\psi)$ from the vacuum $|0\rangle$. On \mathcal{H}_0 we consider the subspace $\mathcal{O}(\mathcal{H}_0)$ that is spanned by the states of the form

$$V^{(N)}(\psi)\chi, \quad N > 0 \quad [18]$$

where $V^{(N)}(\psi)$ is defined by

$$V^{(N)}(\psi) = \sum_{n=0}^h \binom{h}{n} V_{-n-N}(\psi) \quad [19]$$

and h is the conformal weight of ψ . Zhu's algebra is then the quotient space

$$A = \mathcal{H}_0 / \mathcal{O}(\mathcal{H}_0) \quad [20]$$

It actually forms an associative algebra, where the algebra structure is defined by

$$\psi \star \chi = V^{(0)}(\psi)\chi \quad [21]$$

This algebra structure can be identified with the action of the “zero-mode algebra” on an arbitrary highest-weight state.

Zhu’s algebra captures much of the structure of the (chiral) conformal field theory: in particular, it was shown by Zhu in 1996 that the irreducible representations of A are in one-to-one correspondence with the representations of the full vertex operator algebra. A conformal field theory is thus rational (in the above, physicists’, sense) if Zhu’s algebra is finite dimensional. (In the mathematics literature, a vertex operator algebra is usually called rational if in addition every positive-energy representation is completely reducible. It has been conjectured that this is equivalent to the condition that Zhu’s algebra is semisimple.)

In practice, the determination of Zhu’s algebra is quite complicated, and it is therefore useful to obtain more easily testable conditions for rationality. One of these is the so-called C_2 condition of Zhu: a vertex operator algebra is C_2 -cofinite if the quotient space $\mathcal{H}_0/\mathcal{O}_2(\mathcal{H}_0)$ is finite dimensional, where $\mathcal{O}_2(\mathcal{H}_0)$ is spanned by the vectors of the form

$$V_{-n-b}(\psi)\chi, \quad n \geq 1 \quad [22]$$

It is easy to show that the C_2 -cofiniteness condition implies that Zhu’s algebra is finite dimensional. Gaberdiel and Neitzke have shown that every C_2 -cofinite vertex operator algebra has a simple spanning set; this observation can, for example, be used to prove that all the fusion rules (see below) of such a theory are finite.

Fusion Rules and Verlinde’s Formula

As explained above, the correlation function of three primary fields is determined up to an overall constant. One important question is whether or not this constant actually vanishes since this determines the possible “couplings” of the theory. This information is encoded in the so-called fusion rules of the theory. More precisely, the fusion rules $N_{ij}^k \in \mathbb{N}_0$ determine the multiplicity with which the representation of the vertex operator algebra labeled by k appears in the operator-product expansion of the two representations labeled by i and j .

In 1988, Verlinde found a remarkable relation between the fusion rules of a vertex operator algebra and the modular transformation properties of its characters. To each irreducible representation

\mathcal{H}_i of a vertex operator algebra, one can define the character

$$\chi_i(\tau) = \text{tr}_{\mathcal{H}_i} \left(q^{L_0 - (c/24)} \right), \quad q = e^{2\pi i \tau} \quad [23]$$

For rational vertex operator algebras (in the mathematical sense) these characters transform under the modular transformation $\tau \mapsto -1/\tau$ as

$$\chi(-1/\tau) = \sum_j S_{ij} \chi_j(\tau) \quad [24]$$

where S_{ij} are constant matrices. Verlinde’s formula then states that, at least for unitary theories,

$$N_{ij}^k = \sum_l \frac{S_{il} S_{jl} S_{kl}^*}{S_{0l}} \quad [25]$$

where the “0” label denotes the vacuum representation. A general argument for this formula has been given by Moore and Seiberg in 1989; very recently, this has been made more precise by Huang.

Modular Invariance and the Conformal Bootstrap

Up to now, we have only considered conformal field theories on the sphere. In order for the theory to be well defined also on higher-genus surfaces, it is believed that the only additional requirement comes from the consistency of the torus amplitudes. In particular, the vacuum torus amplitude must only depend on the equivalence class of tori that is described by the modular parameter $\tau \in \mathbb{H}$, up to the discrete identifications that are generated by the usual action of the modular group $\text{SL}(2, \mathbb{Z})$ on the upper half-plane \mathbb{H} . For the theory with decomposition [8] this requires that the function

$$Z(\tau, \bar{\tau}) = \sum_{ij} M_{ij} \chi_i(\tau) \chi_j(\bar{\tau}) \quad [26]$$

is invariant under the action of $\text{SL}(2, \mathbb{Z})$. This is a very powerful constraint on the multiplicity matrices M_{ij} that has been analyzed for various vertex operator algebras. For example, Cappelli, Itzykson, and Zuber have shown that the modular invariant WZW models corresponding to the group $\text{SU}(2)$ have an A–D–E classification. The case of $\text{SU}(3)$ was solved by Gannon, using the Galois symmetries of these rational conformal field theories.

The condition of modular invariance is relatively easily testable, but it does not, by itself, guarantee that a given space of states \mathcal{H} comes from a consistent conformal field theory. In order to construct a consistent conformal field theory, one needs to solve the conformal bootstrap, that is, one has to determine

all the normalization constants of the correlators so that the resulting set of correlators is local and factorizes appropriately into 3-point correlators (crossing symmetry). This is typically a difficult problem which has only been solved explicitly for rather few theories, for example, the minimal models. Recently, it has been noticed that the conformal bootstrap can be more easily solved for the corresponding boundary conformal field theory. Furthermore, Fuchs, Runkel, and Schweigert have shown that any solution of the boundary problem induces an associated solution for conformal field theory on surfaces without boundary. This construction relies heavily on the relation between 2D conformal field theory and 3D topological field theory (Turaev 1994).

See also: Boundary Conformal Field Theory; Compactification of Superstring Theory; Current Algebra; Knot Theory and Physics; String Field Theory; Superstring Theories; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions.

Further Reading

Di Francesco P, Mathieu P, and Sénéchal D (1997) *Conformal Field Theory*. New York: Springer.

Frenkel I, Lepowski J, and Meurman A (1988) *Vertex Operator Algebras and the Monster*. Boston, MA: Academic Press.

Gaberdiel MR (2000) An introduction to conformal field theory. *Reports on Progress in Physics* 63: 607 (arXiv:hep-th/9910156).

Gannon T (1999) Monstrous moonshine and the classification of CFT, arXiv:math.QA/9906167.

Gannon T (2006) *Moonshine beyond the Monster: The Bridge Connecting Algebra, Modular Forms and Physics* (to appear). Cambridge: Cambridge University Press.

Gawedzki K (1999) Lectures on conformal field theory. In: *Quantum Fields and Strings: A Course for Mathematicians*, vol. 2. Providence, RI: American Mathematical Society.

Ginsparg P (1988) *Applied Conformal Field Theory*. Lectures Given at the Les Houches Summer School in Theoretical Physics. Elsevier.

Goddard P (1989) Meromorphic conformal field theory. *Infinite Dimensional Lie Algebras and Lie Groups: Proceedings of the CIRM Luminy Conference, 1988*, 556. Singapore: World Scientific.

Goddard P and Olive DI (1988) *Kac-Moody and Virasoro Algebras, A Reprint Volume for Physicists*. Singapore: World Scientific.

Kac VG (1998) *Vertex Algebras for Beginners*. Providence, RI: American Mathematical Society.

Pressley A and Segal GB (1986) *Loop Groups*. Oxford: Clarendon.

Schweigert C, Fuchs J, and Walcher J (2000) Conformal field theory, boundary conditions and applications to string theory, arXiv:hep-th/0011109.

Turaev VG (1994) *Quantum Invariants of Knots and 3-Manifolds*. Berlin: de Gruyter.

Two-Dimensional Ising Model

B M McCoy, State University of New York at Stony Brook, Stony Brook, NY, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The Ising model is a model of a classical ferromagnet on a lattice first introduced in 1925 in the one-dimensional case by E Ising. At each lattice site there is a “spin” variable σ , which takes on the values +1 (spin up) and -1 (spin down). The mutual interaction energy of the pair of spins σ_α and $\sigma_{\alpha'}$, where α and α' are nearest neighbors, is $-E(\alpha, \alpha')$ if $\sigma_\alpha = \sigma_{\alpha'}$ and is $E(\alpha, \alpha')$ if $\sigma_\alpha = -\sigma_{\alpha'}$. In addition, the spins can interact with an external magnetic field as $-H\sigma_\alpha$. On a square lattice, where j specifies the row and k specifies the column, the interaction energy for the homogeneous case where $E_v(\alpha, \alpha')$ and $E_h(\alpha, \alpha')$ are independent of the position α, α' may be explicitly written as

$$E(H) = - \sum_{j,k} [E_h \sigma_{j,k} \sigma_{j,k+1} + E_v \sigma_{j,k} \sigma_{j+1,k} + H \sigma_{j,k}] \quad [1]$$

This very simple model [1] has the remarkable property that in two dimensions at $H=0$ many properties of physical interest can be computed exactly. Furthermore, the model has a ferromagnetic phase transition at a critical temperature T_c , at which the specific heat diverges and the magnetic susceptibility diverges to infinity and below which there is a nonzero spontaneous magnetization. In addition, the microscopic correlations between spins can also be exactly computed. These exact calculations are the basis of the modern theory of second-order phase transitions used to analyze real ferromagnets and real fluids near their critical points in both two and three dimensions. The model may also be interpreted as a lattice gauge theory.

Solvability

The solvability of the Ising model at $H=0$ was discovered by Onsager in 1944 in one of the most profound and inventive papers ever written in mathematical physics. Onsager discovered that the model possesses an infinite-dimensional symmetry, which allowed him to exactly compute the free

energy per site. This symmetry is generated by the relations

$$\begin{aligned} [A_l, A_m] &= 4G_{l-m} \\ [G_l, A_m] &= 2A_{m+l} - 2A_{m-l} \\ [G_l, G_m] &= 0 \end{aligned} \tag{2}$$

This algebra of Onsager is a subalgebra of what is now called the loop algebra of the Lie algebra Sl_2 and it is the first infinite-dimensional algebra to be used in physics.

In the 60 years since Onsager first computed the free energy, several other methods of exact solution have been found. In 1949, Kaufman reduced the computation of the free energy to a problem of free fermions. A closely related combinatorial method was invented by Kac and Ward, Hurst and Green, and by Kastelyn. Baxter (1982) has computed the free energy by means of star triangle equations and functional equations in his book.

The fermionic and the combinatorial methods are powerful enough to compute the correlation functions but are not generalizable to other models. The functional equation methods of Baxter generalize to many other important models but they do not give correlation functions. There are still aspects of Onsager's method that remain unexplored.

The free energy per site in the thermodynamic limit is defined as

$$F = -k_B T \lim_{N \rightarrow \infty} \mathcal{N}^{-1} \ln Z_I(H) \tag{3}$$

where \mathcal{N} is the total number of sites of the lattice and the partition function $Z_I(H)$ is defined as

$$Z_I(H) = \sum_{\text{all } \sigma = \pm 1} e^{E(H)/k_B T} \tag{4}$$

with the sum being over all values $\sigma_{j,k} = \pm 1$ and k_B is Boltzmann's constant. The result of Onsager is that, at $H = 0$,

$$\begin{aligned} F/k_B T &= \ln 2 + \frac{1}{8\pi^2} \int_0^{2\pi} d\theta_1 \int_0^{2\pi} d\theta_2 \ln \left[\cosh 2E_h/k_B T \right. \\ &\quad \times \cosh 2E_v/k_B T - \sinh 2E_h/k_B T \cos \theta_1 \\ &\quad \left. - \sinh 2E_v/k_B T \cos \theta_2 \right] \end{aligned} \tag{5}$$

This free energy has a singularity at a temperature T_c defined from

$$\sinh(2E_v/k_B T_c) \sinh(2E_h/k_B T_c) = 1 \tag{6}$$

and near T_c the specific heat diverges as

$$\begin{aligned} c \sim & -\frac{2}{k_B T^2 \pi} \left(E_h^2 \sinh^2 2E_v/k_B T_c + 2E_v E_h \right. \\ & \left. + E_v^2 \sinh^2 2E_h/k_B T_c \right) \ln |1 - T/T_c| \end{aligned} \tag{7}$$

The next property to be computed was the spontaneous magnetization, which is usually defined as

$$M_- = \lim_{H \rightarrow 0^+} M(H) \tag{8}$$

However, because solution is only available at $H = 0$, this definition cannot be used and instead M_- is computed from an alternative definition in terms of the spin-spin correlation function

$$\langle \sigma_{0,0} \sigma_{M,N} \rangle = \frac{1}{Z_I(0)} \sum_{\sigma = \pm 1} \sigma_{0,0} \sigma_{M,N} e^{-E(0)/k_B T} \tag{9}$$

as

$$M_-^2 = \lim_{M^2 + N^2 \rightarrow \infty} \langle \sigma_{0,0} \sigma_{M,N} \rangle \tag{10}$$

The result for M_- , first announced by Onsager in 1949, is

$$M_- = \begin{cases} (1 - k^2)^{1/8} & \text{for } T \leq T_c \\ 0 & \text{for } T > T_c \end{cases} \tag{11}$$

where

$$k = (\sinh 2E_v/k_B T \sinh 2E_h/k_B T)^{-1} \tag{12}$$

A key point in the computation of the magnetization [11] from [9] is that the spin-spin correlation function can be written as a determinant. In fact, there are many such different, but equal, determinantal representations and the size of the smallest one in general is $2(|M| + |N|)$. The simplest case is the diagonal correlation

$$\langle \sigma_{0,0} \sigma_{N,N} \rangle = \begin{vmatrix} a_0 & a_{-1} & a_{-2} & \cdots & a_{1-N} \\ a_1 & a_0 & a_{-1} & \cdots & a_{2-N} \\ a_2 & a_1 & a_0 & \cdots & a_{3-N} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{N-1} & a_{N-2} & a_{N-3} & \cdots & a_0 \end{vmatrix} \tag{13}$$

where

$$a_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\theta e^{-in\theta} \left[\frac{1 - ke^{-i\theta}}{1 - ke^{i\theta}} \right]^{1/2} \tag{14}$$

Determinants of the form [13], where the elements on each diagonal are equal, are called Toeplitz.

The study of the spin-spin correlations of the Ising model provides a microscopic picture of the behavior of the ferromagnet near the phase transition temperature T_c , and an entire branch of mathematics has developed from the study of the behavior of Toeplitz determinants when the size is large. The first such mathematical advance was the discovery by Szegő of a general formula for the limit as $N \rightarrow \infty$, from which the magnetization [11] is computed.

The simplest result for the approach to this $N \rightarrow \infty$ limit is the behavior of the diagonal correlation at $T = T_c(k = 1)$, where [13] exactly reduces to

$$\langle \sigma_{0,0} \sigma_{N,N} \rangle = \left(\frac{2}{\pi} \right)^N \prod_{l=1}^{N-1} \left[1 - \frac{1}{4l^2} \right]^{l-N} \quad [15]$$

which behaves as $N \rightarrow \infty$ as

$$\langle \sigma_{0,0} \sigma_{N,N} \rangle \sim AN^{-1/4} \quad [16]$$

where $A \sim 0.6450 \dots$ is a transcendental constant. Further results for large N and T fixed are for $T < T_c(k < 1)$,

$$\langle \sigma_{0,0} \sigma_{N,N} \rangle \sim M_-^2 \left\{ 1 + \frac{2k^{2(N+1)}}{\pi N^2(k^{-2} - 1)^2} + \dots \right\} \quad [17]$$

and for $T > T_c(k > 1)$,

$$\langle \sigma_{0,0} \sigma_{N,N} \rangle \sim \frac{k^{-N}}{(\pi N)^{1/2}(1 - k^{-2})^{1/4}} + \dots \quad [18]$$

By comparing [16] with [17] and [18], we see that at $T = T_c$ the correlations decay algebraically but for $T \neq T_c$ the decay is exponential. It is useful to write the exponential in [17] for $T < T_c$ as

$$k^{-N} = e^{-N/\xi_-} \quad \text{with } \xi^{-1} = -\ln k \quad [19]$$

and in [18] for $T > T_c$ as

$$k^N = e^{-N/\xi_+} \quad \text{with } \xi^{-1} = \ln k \quad [20]$$

The quantity ξ is called the correlation length and as $T \rightarrow T_c$ the correlation length diverges as

$$\xi_{\pm} \sim |1 - k|^{-1} = \text{const.} |T - T_c|^{-1} \quad [21]$$

A more profound property of the correlations is that they satisfy differential and difference equations. It was found by Jimbo and Miwa (1980) that the diagonal correlation function satisfies the nonlinear differential equation related to the sixth Painlevé function

$$\begin{aligned} & \left(t(t-1) \frac{d^2 \sigma}{dt^2} \right)^2 \\ &= N^2 \left((t-1) \frac{d\sigma}{dt} - \sigma \right)^2 \\ & - 4 \frac{d\sigma}{dt} \left((t-1) \frac{d\sigma}{dt} - \sigma - \frac{1}{4} \right) \left(t \frac{d\sigma}{dt} - \sigma \right) \end{aligned} \quad [22]$$

where for $T < T_c$ we set $t = k^{-2}$ and

$$\sigma_N(t) = t(t-1) \frac{d}{dt} \ln \langle \sigma_{0,0} \sigma_{N,N} \rangle - \frac{1}{4} \quad [23]$$

and for $T > T_c$ we set $t = k^2$ and

$$\sigma_N(t) = t(t-1) \frac{d}{dt} \ln \langle \sigma_{0,0} \sigma_{N,N} \rangle - \frac{t}{4} \quad [24]$$

Furthermore it was found by McCoy *et al.* (1981) that for a given temperature the general two spin correlation function and all multipoint correlations satisfy quadratic nonlinear partial difference equations in the locations of the spins.

Scaling Theory

It is evident that the results [17] and [18] do not reduce to [16] when $k \rightarrow 1$. Therefore, in order to uniformly characterize the behavior of the correlation function in the critical region near T_c , it is necessary to introduce what is called the scaling function. This uniform expansion is obtained by introducing a scaled length defined as

$$r = N/\xi \quad [25]$$

and considering the joint (scaling limit) where

$$N \rightarrow \infty \quad \text{and} \quad T \rightarrow \infty \quad \text{with } r \text{ fixed} \quad [26]$$

We define the scaled correlation function as

$$G_{\pm}(r) = \lim_{\text{scaling}} M_{\pm}^{-2} \langle \sigma_{0,0} \sigma_{N,N} \rangle \quad [27]$$

where the subscript \pm means that the limit is taken from $T > T_c$ or $T < T_c$, respectively, M_{\pm} is the spontaneous magnetization [11] and

$$M_{+} = (k^{-2} - 1)^{1/8} \quad [28]$$

This concept of the scaling limit and scaling function is very general and can be defined for any system with a critical point that has an order parameter like M_{\pm} that vanishes at T_c and a correlation length that diverges at T_c . However, the Ising model has the further remarkable property discovered by Wu *et al.* (1976) that the scaled correlation function may be explicitly expressed in terms of a function which satisfies an ordinary nonlinear differential equation. Specifically,

$$\begin{aligned} G_{\pm}(r) &= \frac{1}{2} [1 \mp \eta(r/2)] \eta(r/2)^{-1/2} \\ & \times \exp \int_{r/2}^{\infty} r' \frac{r'}{4} \eta^{-2} [(1 - \eta^2)^2 - (\eta')^2] \end{aligned} \quad [29]$$

where the function $\eta(r)$ satisfies the Painlevé III equation

$$\eta'' = \frac{1}{\eta} (\eta')^2 - \frac{\eta'}{r} + \eta^3 - \eta^{-1} \quad [30]$$

with the boundary condition that

$$\eta(r) \sim 1 - 2\lambda K_0(2r) \quad \text{as } r \rightarrow \infty \quad [31]$$

where $K_0(r)$ is the modified Bessel function of the third kind and

$$\lambda = 1/\pi \quad [32]$$

The leading behavior of $G_{\pm}(r)$ for $r \rightarrow \infty$ is

$$G_+(r) \sim \lambda K_0(r) \quad [33]$$

$$G_-(r) \sim 1 + \lambda^2 \left\{ r^2 [K_1^2(r) - K_0^2(r)] - r K_0(r) K_1(r) + \frac{1}{2} K_0^2(r) \right\} \quad [34]$$

where $K_n(z)$ is the modified Bessel function of the third kind. When λ is given by [32] these $r \rightarrow \infty$ limits of $G_{\pm}(r)$ agree with the behavior of $\langle \sigma_{0,0} \sigma_{N,N} \rangle$ for $N \gg 1$ and $|T - T_c|$ small with $N|T - T_c| \gg 1$ which is obtained from [18] and [17]. The behavior of $G_{\pm}(r)$ for $r \rightarrow 0$ with the value of λ given by [32] is

$$G_{\pm}(r) = \text{const. } r^{-1/4} \quad [35]$$

where the constant agrees with that computed from the result [16] for $\langle \sigma_{0,0} \sigma_{N,N} \rangle$ at $T = T_c$ for $N \gg 1$. For other values of the boundary condition constant λ , the scaling function $G_{\pm}(r)$ diverges with a power which differs from $1/4$. The computation of the constant in [35] requires the evaluation of a nontrivial integral involving the Painlevé III function.

The agreement of the limits $r \rightarrow \infty$ and $r \rightarrow 0$ of the function $G_{\pm}(r)$ with the lattice results near T_c means that this scaling function uniformly interpolates between $T \neq T_c$ and $T = T_c$ and that the lattice size (defined here as unity) and the self-generated correlation length ξ are the only two length scales in the theory. This feature that the system generates only one new length scale near T_c is referred to as one length scale scaling.

Susceptibility

The final quantity of macroscopic thermodynamic interest is the magnetic susceptibility

$$\chi(T) = \left. \frac{\partial M(H)}{\partial H} \right|_{H=0} \quad [36]$$

which is expressed in terms of the spin-spin correlation function as

$$\chi(T) = \frac{1}{k_B T} \sum_{M,N} \{ \langle \sigma_{0,0} \sigma_{M,N} \rangle - M_-^2 \} \quad [37]$$

The susceptibility may be studied by using the determinantal expression for the correlation function. The simplest result is obtained (for the isotropic case, $E_v = E_h$) by using the scaling form [27] to find for $T \sim T_c$ that

$$k_B T \chi_{\pm}(T) \sim M_{\pm}^2 \xi^2 2\pi \int_0^{\infty} dr r \{ G_{\pm} - \Theta_{\pm} \} \quad [38]$$

where $\Theta_+ = 0$ and $\Theta_- = 1$, and thus $\chi_{\pm}(T)$ diverges at $T \rightarrow T_c$ as

$$\chi_{\pm}(T) \sim C_{\pm} |T - T_c|^{-7/4} \quad [39]$$

where C_{\pm} are transcendental constants given as integrals over the scaling function $G_{\pm}(r)$, which were first evaluated by Barouch *et al.* in 1973 as

$$\begin{aligned} C_- &= 0.0255369719 \dots, \\ C_+ &= 0.9625817322 \dots \end{aligned} \quad [40]$$

Critical-Exponent Phenomenology

From the behavior for the Ising model of the specific heat, magnetization, susceptibility, correlation length, and the correlation at T_c given above we abstract for general systems the phenomenological critical-exponent parametrization for $T \rightarrow T_{c\pm}$ of

$$c \sim A_c^{\pm} |T - T_c|^{-\alpha_{\pm}} \quad [41]$$

$$M \sim A_M |T_c - T|^{\beta} \quad [42]$$

$$\chi \sim A_{\chi}^{\pm} |T - T_c|^{-\gamma_{\pm}} \quad [43]$$

$$\xi \sim A_{\xi}^{\pm} |T - T_c|^{-\nu_{\pm}} \quad [44]$$

and at $T = T_c$ for $R \rightarrow \infty$

$$\langle \sigma_0 \sigma_R \rangle \sim A_{\sigma} / R^{d-2+\eta} \quad \text{where } d \text{ is the dimension} \quad [45]$$

The exponents $\alpha_{\pm}, \gamma_{\pm}, \nu_{\pm}$ above and below T_c are usually found to be equal, and the exponent η is usually called the anomalous dimension. If it is assumed that the scaling function [27] exists and that one length scale scaling holds then the exponents are related by what are called scaling laws, such as

$$2\beta = \nu_-(d - 2 + \eta) \quad [46]$$

$$\alpha_- + 2\beta - \gamma_- = 2 \quad [47]$$

$$d\nu_- = 2 - \alpha_- \quad [48]$$

Thus, from the properties of the Ising model near T_c , we have obtained a phenomenology for use on all systems near the critical point.

Fuchsian Equations and Natural Boundaries for Susceptibility

This critical phenomenology, however, has not taken into account the fact that the susceptibility is a much more complicated function than either the spontaneous magnetization [11] or the free energy [5], which have only isolated singularities at $k^2=1$, and that there is more structure to the susceptibility than the singularity of [39].

For arbitrary T , the susceptibility was shown by Wu, McCoy, Tracy, and Barouch to be expressible in the form

$$\chi_{\pm}(T) = M_{\pm}^2 \sum_j \tilde{\chi}^{(j)}(T) \quad [49]$$

where in the sum j is odd (even) for T above (below) T_c . The quantities $\tilde{\chi}^{(j)}(T)$ are explicitly given as j -fold integrals of algebraic functions and thus will satisfy linear differential equations with polynomial coefficients. Such functions can have only isolated singularities. The function $\tilde{\chi}^{(1)}(T)$ is elementary and has a double pole at T_c and $\tilde{\chi}^{(2)}(T)$ is given in terms of complete elliptic integrals. Quite recently, remarkable Fuchsian linear differential equations for $\tilde{\chi}^{(3)}(T)$ and $\tilde{\chi}^{(4)}(T)$ of seventh and tenth orders, respectively, have been obtained by Zenine, Bouk-raa, Hassani, and Maillard for the isotropic lattice.

Furthermore, it was shown by Orrick *et al.* (2001) that $\tilde{\chi}^{(j)}$ has singularities in the complex T plane at

$$\begin{aligned} & \cosh(2E_v/kT) \cosh(2E_h/kT) \\ & - \sinh(2E_v/kT) \cos(2\pi/j) \\ & - \sinh(2E_h/kT) \cos(2\pi m'/j) = 0 \end{aligned} \quad [50]$$

with $m, m' = 1, 2, \dots, j$. The form of the singularity in $\tilde{\chi}^{(j)}(T)$ for $T > T_c$ is as

$$\epsilon^{(j^2-3)/2} \ln \epsilon \quad [51]$$

and, for $T < T_c$, it is as

$$\epsilon^{(j^2-3)/2} \quad [52]$$

where ϵ measures the deviation from the singular point [50]. These singularities become dense as $j \rightarrow \infty$ and, therefore, the singularity at $T=T_c$ is not isolated and instead the critical point is embedded in a natural boundary. Such a function cannot satisfy a linear differential equation of finite order with polynomial coefficients.

The existence of the natural boundary in the susceptibility is a new phenomenon which is not seen in either the free energy or magnetization and leads to the speculation that in the presence of a magnetic field the one length scale scaling property of the model at $H=0$ may fail. If this proves to be

correct, there will be physical effects which are not incorporated in the phenomenological scaling theory of critical phenomena.

Impure Ising Models

The Ising model may also be studied when the interaction energies at sites j, k are not chosen to be independent of position but are allowed to vary from site to site. When these interactions are chosen randomly out of some probability distribution, this is a model of a ferromagnet with frozen (quenched) impurities. All real systems will be impure to some extent, so the study of such dirty systems is of great practical importance.

The special case where the interactions are translationally invariant in the horizontal direction but are allowed to vary in a layered fashion from row to row was introduced by McCoy and Wu in 1968 and found to be dramatically different from the pure Ising model described above. In particular, what is a critical temperature T_c in the pure case is now spread out into a region bounded by the temperatures the pure model would be critical if all the bonds took on the minimum or maximum value allowed by the probability distribution. In this new region, the correlations (in the direction of translational invariance) are found to decay as a power law which depends on the temperature; the specific heat is never infinite but the susceptibility is infinite in an entire temperature region that includes the temperature at which the spontaneous magnetization first appears as T is lowered. The existence of this new region for Ising models with a general randomness in two and three dimensions has been demonstrated by Griffiths. More recently, this effect has been reinterpreted in terms of impurities in quantum spin chains.

Quantum Field Theory

The Ising model of [1] may be reinterpreted as a two-dimensional lattice gauge theory of the gauge field

$$\begin{aligned} & s_{j+1/2, k} = \pm 1 \\ & \quad \text{on the vertical link between } (j, k) \text{ and } (j+1, k) \\ & s_{j, k+1/2} = \pm 1 \\ & \quad \text{on the horizontal link between } (j, k) \\ & \quad \text{and } (j, k+1) \\ & \text{and a "Higgs" field} \\ & \phi_{j, k} = \pm 1 \quad \text{on the site } (j, k) \end{aligned} \quad [53]$$

$$\phi_{j, k} = \pm 1 \quad \text{on the site } (j, k) \quad [54]$$

with the action

$$S_g = -E_g \sum_{j,k} s_{j+1/2,k} s_{j+1,k+1/2} s_{j+1/2,k+1} s_{j,k+1/2} - E_h \sum_{j,k} (\phi_{j,k} s_{j+1/2,k} \phi_{j+1,k} + \phi_{j,k} s_{j,k+1/2} \phi_{j,k+1}) \quad [55]$$

If we define

$$z_{g,h} = \tanh E_{g,h}/k_B T \quad [56]$$

the partition function of the gauge theory is expressed in terms of the Ising model partition function as

$$Z_g = [8 \cosh(E_g/k_B T) \cosh^2(E_h/k_B T) z_g^{1/2} z_h]^N Z_I(H) \quad [57]$$

where we make the identification

$$H/k_B T = \frac{1}{2} \ln z_g \quad \text{and} \quad E/k_B T = \frac{1}{2} \ln z_h \quad [58]$$

This identification may be extended to correlation functions. Of particular interest for the gauge theory is the plaquette-plaquette correlation $\langle P_{0,0} P_{j,l} \rangle$, where

$$P_{j,k} = s_{j+1/2,k} s_{j+1,k+1/2} s_{j+1/2,k+1} s_{j,k+1/2} \quad [59]$$

which is expressed in terms of the Ising correlations at $H \neq 0$ as

$$\begin{aligned} \langle P_{0,0} P_{j,k} \rangle &= -\langle P_{0,0} \rangle^2 \\ &= \sinh^2(2H/k_B T) (\langle \sigma_{0,0} \sigma_{j,k} \rangle - \langle \sigma_{0,0} \rangle^2) \end{aligned} \quad [60]$$

To study this correlation further, we need to study the correlations of the Ising model in nonzero magnetic field. This has been done by McCoy and Wu in the scaling limit $H \rightarrow 0, T \rightarrow T_c$ with

$$h = \frac{H}{|T - T_c|^{15/8}} \text{ fixed} \quad [61]$$

for $T < T_c$, where it is found that the scaling function $G(r, h)$ for small h and large r if

$$\begin{aligned} G(r, h) &\sim \sum_l a b K_0 \left[(2 + h^{2/3} \lambda_l) r \right] \\ &\sim \pi^{1/2} r^{-1/2} e^{-2r} \sum_l h a e^{-r h^{2/3} \lambda_l} \end{aligned} \quad [62]$$

where λ_l are the solutions of

$$J_{1/3}(\frac{1}{3} \lambda^{3/2}) + J_{-1/3}(\frac{1}{3} \lambda^{3/2}) = 0 \quad [63]$$

with $J_n(z)$ the Bessel function of order n and $K_0(z)$ the modified Bessel function of the third kind.

A field theory is said to possess a particle spectrum if the Fourier transform of the two-point function

$$G(k, h) = \int d^2 r e^{i k \cdot r} G(r, h) \quad [64]$$

has poles of the form $A_l/(k^2 + m_l^2)$, where m_l is the mass of the l th particle. If we note that the Fourier transform of $K_0(r)$ is

$$\int d^2 r e^{i k \cdot r} K_0(r) = \frac{2\pi}{k^2 + 1} \quad [65]$$

we see that the Fourier transform of [62] is the sum of an infinite number of poles. This is to be compared with the Fourier transform of the scaled correlation function $G_-(r)$ at $H = 0$ and $T < T_c$ [34], which does not contain any poles at all and may instead be interpreted as having a two-particle cut. This phenomenon of a cut at $h = 0$ breaking up into an infinite number of poles for $h > 0$ is a signal that at $h = 0$ the theory has free unconfined two-particle states which become weakly confined by a linear confining potential for $h > 0$. This confinement is thought to be a characteristic of most gauge theories.

See also: Eight Vertex and Hard Hexagon Models; Holonomic Quantum Fields; Painlevé Equations; Percolation Theory; Phase Transitions in Continuous Systems; Statistical Mechanics and Combinatorial Problems; Toeplitz Determinants and Statistical Mechanics; Yang-Baxter Equations.

Further Reading

- Barouch E, McCoy BM, and Wu TT (1973) Zero-field susceptibility of the two dimensional Ising model near T_c . *Physical Review Letters* 31: 1409–1411.
- Baxter RJ (1982) *Exactly Solved Models in Statistical Mechanics*. London: Academic Press.
- Griffiths RB (1969) Nonanalytic behavior above the critical point in a random Ising ferromagnet. *Physical Review Letters* 23: 17–19.
- Jimbo M and Miwa T (1980) Studies on holonomic quantum fields. XVII. *Proceedings of the Japanese Academy* 56A: 405; (1981) 57A: 347.
- Kasteleyn PW (1963) Dimer statistics and phase transitions. *Journal of Mathematical Physics* 4: 287–293.
- McCoy BM (1969) Incompleteness of the critical exponent description for ferromagnetic systems containing random impurities. *Physical Review Letters* 23: 383–388.
- McCoy BM (1995) The connection between statistical mechanics and quantum field theory. In: Bazhanov VV and Burden CJ (eds.) *Statistical Mechanics and Field Theory*, pp. 26–128. Singapore: World Scientific.
- McCoy BM and Wu TT (1968) Theory of a two dimensional Ising model with random impurities. *Physical Review* 176: 631–643.
- McCoy BM and Wu TT (1973) *The Two Dimensional Ising Model*. Cambridge: Harvard University Press.
- McCoy BM and Wu TT (1978) Two dimensional Ising field theory in a magnetic field: Breakup of the cut in the 2-point function. *Physical Review D* 18: 1259–1267.
- McCoy BM, Perk JHH, and Wu TT (1981) Ising field theory: quadratic difference equations for the n -point Green's functions on the lattice. *Physical Review Letters* 46: 757.
- Montroll EW, Potts RB, and Ward JC (1963) Correlations and spontaneous magnetization of the two dimensional Ising model. *Journal of Mathematical Physics* 4: 308–322.

- Onsager L (1944) Crystal statistics I. A two dimensional model with an order disorder transition. *Physical Review* 65: 117–149.
- Orrick WP, Nickel BG, Guttman AJ, and Perk JHH (2001) Critical behavior of the two dimensional Ising susceptibility. *Physical Review Letters* 86: 4120–4123.
- Wu TT, McCoy BM, Tracy CA, and Barouch E (1976) Spin–spin correlation functions for the two dimensional Ising model, exact theory in the scaling region. *Physical Review B* 13: 315–374.

- Zenine N, Boukraa S, Hassani S, and Maillard JM (2004) The Fuchsian differential equation of the square lattice Ising model $\chi^{(3)}$ susceptibility. *Journal of Physics A* 37: 9651–9668.
- Zenine N, Boukraa S, Hassani S, and Maillard JM (2005) Ising model susceptibility: Fuchsian differential equation for $\chi^{(4)}$ and its factorization properties. *Journal of Physics A* 38: 4149–4173.

Two-Dimensional Models

B Schroer, Freie Universität Berlin, Berlin, Germany

© 2006 Elsevier Ltd. All rights reserved.

History and Motivation

Local quantum physics of systems with infinitely many interacting degrees of freedom leads to situations whose understanding often requires new physical intuition and mathematical concepts beyond that acquired in quantum mechanics and perturbative constructions in quantum field theory. In this situation, two-dimensional soluble models turned out to play an important role. On the one hand, they illustrate new concepts and sometimes remove misconceptions in an area where new physical intuition is still in the process of being formed. On the other hand, rigorously soluble models confirm that the underlying physical postulates are mathematically consistent, a task which for interacting systems with infinite degrees of freedom is mostly beyond the capability of pedestrian methods or brute force application of hard analysis on models whose natural invariances have been mutilated by a cutoff.

In order to underline these points and motivate the interest in two-dimensional QFT, let us briefly look at the history, in particular at the physical significance of the three oldest two-dimensional models of relevance for statistical mechanics and relativistic particle physics, in chronological order: the Lenz–Ising (L–I) model, Jordan’s model of bosonization/fermionization, and the Schwinger model (QED₂). (A more detailed account of the changeable history concerning their correct physical interpretation and generalizations to higher dimensions of these models and the increasing conceptual role of low-dimensional models in QFT can be found in Schroer (2005).)

The L–I model was proposed in 1920 by Wilhelm Lenz (see Lenz (1925)) as the simplest discrete statistical mechanics model with a chance to go beyond the P. Weiss phenomenological ansatz

involving long-range forces and instead explain ferromagnetism in terms of nonmagnetic short-range interactions. Its one-dimensional version was solved four years later by his student Ernst Ising. Its changeable history reached a temporary conceptual climax when Onsager succeeded to rigorously establish a second-order phase transition in two dimensions.

Another conceptually rich model which lay dormant for almost two decades as a result of a misleading speculative higher-dimensional generalization by its protagonist is the bosonization/fermionization model first proposed by Jordan (1937). This model establishes a certain equivalence between massless two-dimensional fermions and bosons and is related to Thirring’s massless 4-fermion coupling model and also to Luttinger’s one-dimensional model of an electron gas (Schroer). One reason why even nowadays hardly anybody knows Jordan’s contribution is certainly the ambitious but unfortunate title “the neutrino theory of light” under which he published a series of papers.

Both discoveries demonstrate the usefulness of having controllable low-dimensional models; at the same time, their complicated history also illustrates the danger of rushing to premature “intuitive” conclusions about extensions to higher dimensions.

A review of the early historical benchmarks of conceptual progress through the study of solvable two-dimensional models would be incomplete without mentioning Schwinger’s (1962) proposed solution of two-dimensional quantum electrodynamics, afterwards referred to as the Schwinger model. He used this model in order to argue that gauge theories are not necessarily tied to zero-mass vector particles. Some work was necessary (Schroer) to unravel its physical content with the result that the would-be charge of that QED₂ model was “screened” and its apparent chiral symmetry broken; in other words, the model exists only in the so-called Schwinger–Higgs phase with massive free scalar particles accounting for its physical content. Another closely related aspect of

this model which also arose in the Lagrangian setting of four-dimensional gauge theories was that of the θ -angle parametrizing, an ambiguity in the quantization.

A coherent and systematic attempt at a mathematical control of two-dimensional models came in the wake of Wightman's first rigorous programmatic formulation of QFT (Schroer 2005). This formulation stayed close to the physical ideas underlying the impressive success of renormalized QED perturbation theory, although it avoided the direct use of Lagrangian quantization. The early attempts towards a "constructive QFT" found their successful realization in two-dimensional QFT (the $P\varphi_2$ models (Glimm and Jaffe 1987)); the restriction to low dimensions is related to the mild short-distance singularity behavior (super-renormalizability) which these methods require. We will focus our main attention on alternative constructive methods which, even though not suffering from such short-distance restrictions, also suffer from a lack of mathematical control in higher spacetime dimensions; the illustration of the constructive power of these new methods comes presently from massless $d=1+1$ conformal and chiral QFT as well as from massive factorizing models.

There are several books and review articles (Furlan *et al.* 1989, Ginsparg 1990, Di Francesco *et al.* 1996) on $d=1+1$ conformal as well as on massive factorizing models (Abdalla *et al.* 1991). To the extent that concepts and mathematical structures are used which permit no extension to higher dimensions (Kac–Moody algebras, loop groups, integrability, presence of an infinite number of conservation laws), this line of approach will not be followed in this article since our primary interest will be the use of two-dimensional models of QFT as "theoretical laboratories" of general QFT. Our aim is twofold; on the one hand, we intend to illustrate known principles of general QFT in a mathematically controllable context and on the other hand, we want to identify new concepts whose adaptation to QFT in $d=1+1$ lead to their solvability (Schroer).

General Concepts and Their Two-Dimensional Manifestation

The general framework of QFT, to which the rich world of controllable two-dimensional models contributes as an important testing ground, exists in two quite different but nevertheless closely related formulations: the 1956 approach in terms of pointlike covariant fields due to Wightman (see Streater

and Wightman (1964)) (see Axiomatic Quantum Field Theory), and the more algebraic setting which can be traced back to ideas which Haag (1992) developed shortly after and which are based on spacetime-indexed operator algebras and related concepts which developed over a long period of time, with contributions of many other authors to what is now referred to as algebraic QFT (AQFT) or simply local quantum physics (LQP). Whereas the Wightman approach aims directly at the (not necessarily observable) quantum fields, the operator-algebraic setting (see Algebraic Approach to Quantum Field Theory) is more ambitious. It starts from physically well-motivated assumptions about the algebraic structure of local observables and aims at the reconstruction of the full field theory (including the operators carrying the superselected charges) in the spirit of a local representation theory of (the assumed structure of the) local observables. This has the advantage that the somewhat mysterious concept of an inner symmetry (as opposed to outer (spacetime) symmetry) can be traced back to its physical roots which is the representation-theoretical structure of the local observable algebra (see Symmetries in Quantum Field Theory of Lower Spacetime Dimensions). In the standard Lagrangian quantization approach, the inner symmetry is part of the input (multiplicity indices of field components on which subgroups of $U(n)$ or $O(n)$ act linearly) and hence it is not possible to problematize this fundamental question. When in low-dimensional spacetime dimensions the sharp separation (the Coleman–Mandula theorems) of inner versus outer symmetry becomes blurred as a result of the appearance of braid group statistics, the standard Lagrangian quantization setting of most of the textbooks is inappropriate and even the Wightman framework has to be extended. In that case, the algebraic approach is the most appropriate.

The important physical principles which are shared between the Wightman approach (see Streater and Wightman (1964)) and the operator algebra (AQFT) setting (Haag 1992) are the spacelike locality or Einstein causality (in terms of pointlike fields or algebras localized in causally disjoint regions) and the existence of positive-energy representations of the Poincaré group implementing covariance and the stability of matter. In the algebraic approach, the observable content of the theory is encoded into a family of (weakly closed) operator algebras $\{\mathcal{A}(\mathcal{O})\}_{\mathcal{O} \in \mathcal{K}}$ indexed by a family of convex causally closed spacetime regions \mathcal{O} (with \mathcal{O}' denoting the spacelike complement and \mathcal{A}' the von Neumann commutant) which act in one common Hilbert space. Covariant local fields lose their distinguished role

which they have in the classical setting and which (via Lagrangian quantization) was at least partially inherited by the Wightman approach and, apart in their role as local generators of symmetries (conserved currents), became mere “field coordinatizations” of local algebras. (There is a denumerable set of such pointlike field generators which form a local equivalence (Borchers) class of fields and in the absence of interactions permits a neat description in terms of Wick-ordered free-field polynomials (Haag 1992). Certain properties cannot be naturally formulated in the pointlike field setting (e.g., Haag duality for convex regions $\mathcal{A}(\mathcal{O}') = \mathcal{A}(\mathcal{O})'$), but apart from those properties the two formulations are quite close; in particular for two-dimensional theories there are convincing arguments that one can pass between the two without imposing additional technical requirements. (Haag duality holds for observable algebras in the vacuum sector in the sense that any violation can be explained in terms of a spontaneously broken symmetry; in local theories, it can always be enforced by dualization and the resulting Haag dual algebra has a charge superselection structure associated with the unbroken subgroup.) Haag duality is the statement that the commutant of observables not only contains the algebra of the causal complement that is, $\mathcal{A}(\mathcal{O}') \subset \mathcal{A}(\mathcal{O})'$ (Einstein causality) but is even exhausted by it; it is deeply connected to the measurement process and its violation in the vacuum sector for convex causally complete regions signals spontaneous symmetry breaking in the associated charge-carrying field algebra (Haag 1992). It can always be enforced (assuming that the wedge-localized algebras fulfill [1] below) by symmetry-reducing extension called Haag dualization. Its violation for multilocal region reveals the charge content of the model via charge-anticharge splitting in the neutral observable algebra (Schroer).

Another physically important property which has a natural algebraic formulation is the split property: for regions \mathcal{O}_i separated by a finite spacelike distance, one finds $\mathcal{A}(\mathcal{O}_1 \cup \mathcal{O}_2) \simeq \mathcal{A}(\mathcal{O}_1) \otimes \mathcal{A}(\mathcal{O}_2)$ which can be derived from the Buchholz–Wichmann “nuclearity property” (Haag 1992) (an appropriate adaptation of the “finiteness of phase-space cell” property of QM to QFT). Related to the Haag duality is the local version of the “time slice property” (the QFT counterpart of the classical causal dependency property) sometimes referred to as “strong Einstein causality” $\mathcal{A}(\mathcal{O}'') = \mathcal{A}(\mathcal{O})''$.

One of the most astonishing achievements of the algebraic approach (which justifies its emphasis on properties of “local observables”) is the DHR theory of superselection sectors (Doplicher *et al.* 1971),

that is, the realization that the structure of charged (nonvacuum) representations (with the superposition principle being valid only within one representation) and the spacetime properties of the generating fields which are the carriers of these generalized charges (including their spacelike commutation relations which lead to the particle statistics and also to their internal symmetry properties) are already encoded in the structure of the Einstein causal observable algebra (Symmetries in Quantum Field Theory: Algebraic Aspects). The intuitive basis of this remarkable result (whose prerequisite is locality) is that one can generate charged sectors by spatially separating charges in the vacuum (neutral) sector and disposing of the unwanted charges at spatial infinity (Haag 1992).

An important concept which especially in $d = 1 + 1$ has considerable constructive clout is “modular localization.” It is a consequence of the above algebraic setting if either the net of algebras have pointlike field generators, or if the one-particle masses are separated by spectral gaps so that the formalism of time-dependent scattering can be applied (Schroer 2005); in conformal theories, this property holds automatically in all spacetime dimensions. It rests on the basic observation (Tomita–Takesaki Modular Theory) that a standard pair (\mathcal{A}, Ω) of a von Neumann operator algebra and a standard vector (standardness means that the operator algebra of the pair (\mathcal{A}, Ω) acts cyclic and separating on the vector Ω) gives rise to a Tomita operator S through its star-operation whose polar decomposition yield two modular objects, a one-parametric subgroup Δ^{it} of the unitary group of operators in Hilbert space whose Ad-action defines the modular automorphism of (\mathcal{A}, Ω) whereas the angular part J is the modular conjugation which maps \mathcal{A} into its commutant \mathcal{A}'

$$SA\Omega = A^*\Omega, \quad S = J\Delta^{1/2}$$

$$J_W = U(j_W) = S_{\text{scat}} J_0, \quad \Delta_W^{it} = U(\Lambda_W(2\pi t)) \quad [1]$$

$$\sigma_W(t) := \text{Ad} \Delta_W^{it}$$

The standardness assumption is always satisfied for any field-theoretic pair $(\mathcal{A}(\mathcal{O}), \Omega)$ of a \mathcal{O} -localized algebra and the vacuum state (as long as \mathcal{O} has a nontrivial causal disjoint \mathcal{O}'), but it is only for the wedge region W that the modular objects have a physical interpretation in terms of the global symmetry group of the vacuum as specified in the second line of [1]; the modular unitary Δ_W^{it} represents the W -associated boost $\Lambda_W(\chi)$ and the modular conjugation J_W implements the TCP-like reflection along the edge of the wedge (Bisognano

and Wichmann 1975). The third line is the definition of the modular group. The importance of this theory for local quantum physics results from the fact that it leads to the concept of modular localization, an intrinsic new scenario for field-theoretic constructions which is different from the Lagrangian quantization schemes (Schroer 2005).

A special feature of $d = 1 + 1$ Minkowski spacetime is the disconnectedness of the right/left spacelike region leading to a right-left ordering structure. So in addition to the Lorentz-invariant timelike ordering $x \prec y$ (x earlier than y , which is independent of spacetime dimensions), there is an invariant spacelike ordering $x < y$ (x to the left of y) in $d = 1 + 1$ which opens the possibility of more general Lorentz-invariant spacelike commutation relations than those implemented by Bose/Fermi fields (Rehren and Schroer 1987) of fields with a spacelike braid group commutation structure. The appearance of such exotic statistics fields is not compatible with their Fourier transforms being creation/annihilation operators for Wigner particles; rather, the state vectors which they generate from the vacuum contain in addition to the one-particle contribution a vacuum polarization cloud (Schroer 2005). This close connection between new kinematic possibilities and interactions is one of the reasons why, (different from higher dimensions where interactions are prescribed by the recipe of local couplings of free fields) low-dimensional QFT offers a more intrinsic access to the central issue of interactions.

Boson/Fermion Equivalence and Superselection Theory in a Special Model

The simplest and oldest but conceptually still rich model is obtained, as first proposed by Jordan (1937), by using a two-dimensional massless Dirac current and showing that it may be expressed in terms of scalar canonical Bose creation/annihilation operators

$$\begin{aligned} j_\mu &:= \bar{\psi} \gamma_\mu \psi := \partial_\mu \phi, \phi \\ &:= \int_{-\infty}^{+\infty} \{e^{ipx} a^*(p) + h.c.\} \frac{dp}{2|p|} \end{aligned} \quad [2]$$

Although the potential $\phi(x)$ of the current as a result of its infrared divergence is not a field in the standard sense of an operator-valued distribution in the Fock space of the $a(p)^\#$ (It becomes an operator after smearing with test functions whose Fourier transform vanishes at $p=0$), the formal exponential defined as the zero-mass limit of a well-defined exponential free massive field

$$:e^{i\alpha\phi(x)}:= \lim_{m \rightarrow 0} m^{\alpha^2/2} :e^{i\alpha\phi_m(x)}: \quad [3]$$

turns out to be a *bona fide* quantum field in a larger Hilbert space (which extends the Fock space generated from applying currents to the vacuum). The power in front is determined by the requirement that all Wightman functions (computed with the help of free-field Wick combinatorics) stay finite in this massless limit; the necessary and sufficient condition for this is the charge conservation rule

$$\begin{aligned} &\left\langle \prod_i :e^{i\alpha_i\phi(x)}: \right\rangle \\ &= \begin{cases} \prod_{i < j} \left(\frac{-1}{(\xi_{+ij})_\varepsilon (\xi_{-ij})_\varepsilon} \right)^{(1/2)\alpha_i\alpha_j}, & \Sigma\alpha_i = 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad [4]$$

where the resulting correlation function has been factored in terms of light-ray coordinates $\xi_{\pm ij} = x_{\pm i} - x_{\pm j}$, $x_\pm = t \pm x$, and the ε -prescription stands for taking the standard Wightman boundary value $t \rightarrow t + i\varepsilon$, $\lim_{\varepsilon \rightarrow 0}$ which insures the positive-energy condition. The finiteness of the limit insures that the resulting zero-mass limiting theory is a *bona fide* quantum field theory that is, its system of Wightman functions permits the construction of an operator theory in a Hilbert space with a distinguished vacuum vector.

The factorization into light-ray components [4] shows that the exponential charge-carrying operators inherit this factorization into two independent chiral components $: \exp i\alpha\phi(x) := \exp i\alpha\phi_+(x_+): : \exp i\alpha\phi_-(x_-):$, each one being covariant under scaling $\xi \rightarrow \lambda\xi$ if one assigns the scaling dimension $d = \alpha^2/2$ to the chiral exponential field and $d = 1$ to the current. As any Wightman field, this is a singular object which only after smearing with Schwartz test functions yields an (unbounded) operator. But the above form of the correlation function belongs to a class of distributions which admits a much larger test-function space consisting of smooth functions which instead of decreasing rapidly only need to be bounded so that they stay finite on the compactified light-ray line $\dot{R} = S^1$. To make this visible, one uses the Cayley transform (now x denotes either x_+ or x_-)

$$z = \frac{1 + ix}{1 - ix} \in S^1 \quad [5]$$

This transforms the Schwartz test function into a space of test functions on S^1 which have an infinite order zero at $z = -1$ (corresponding to $x = \pm\infty$) but the rotational transformed fields $j(z), : \exp i\alpha\phi(z):$ permit the smearing with all smooth functions on S^1 , a characteristic feature of all conformal invariant

theories as the present one turns out to be. There is an additional advantage in the use of this compactification. Fourier transforming the circular current actually allows for a quantum-mechanical zero mode whose possible nonzero eigenvalues indicate the presence of additional charge sectors beyond the charge-zero vacuum sector. For the exponential field, this leads to a quantum-mechanical pre-exponential factor which automatically insures the charge selection rules so that unrestricted (by charge conservation) Wick contraction rules can be applied. In this approach, the original chiral Dirac fermion $\psi(x)$ (from which the current was formed as the $:\psi\bar{\psi}:$ composite) reappears as a charge-carrying exponential field for $\alpha=1$ and thus illustrates the meaning of bosonization/fermionization. (It is interesting to note that Jordan’s (1937) original treatment of fermionization had such a pre-exponential quantum-mechanical factor.) Naturally, this terminology has to be taken with a grain of salt in view of the fact that the bosonic current algebra only generates a superselected subspace into which the charge-carrying exponential field does not fit. Only in the case of massive two-dimensional QFT fermions can be incorporated into a Fock space of bosons (see last section). At this point, it should however be clear to the reader that the physical content of Jordan’s paper had nothing to do with its misleading title “neutrino theory of light” but rather was an early illustration about charge superselection rules in two-dimensional QFT.

A systematic and rigorous approach consists in solving the problem of positive-energy representation theory for the Weyl algebra on the circle (which is the rigorous operator-algebraic formulation of the abelian current algebra). (The Weyl algebra originated in quantum mechanics around 1927; its use in QFT only appeared after the cited Jordan paper. By representation we mean here a regular representation in which the exponentials can be differentiated in order to obtain (unbounded) smeared current operators.) It is the operator algebra generated by the exponential of a smeared chiral current (always with real test functions) with the following relation between the generators

$$\begin{aligned} W(f) &= e^{ij(f)} \\ j(f) &= \int \frac{dz}{2\pi i} j(z)f(z), [j(z), j(z')] \\ &= -\delta'(z - z') \end{aligned} \tag{6a}$$

$$\begin{aligned} W(f)W(g) &= e^{-(1/2)s(f,g)}W(f+g) \\ W^*(f) &= W(-f) \end{aligned} \tag{6b}$$

$$\begin{aligned} \mathcal{A}(S^1) &= \text{alg}\{W(f), f \in C_\infty(S^1)\} \\ \mathcal{A}(I) &= \text{alg}\{W(f), \text{supp} f \subset I\} \end{aligned} \tag{6c}$$

where

$$s(\cdot, \cdot) = \int \frac{dz}{2\pi i} f'(z)g(z)$$

is the symplectic form which characterizes the Weyl algebra structure and [6c] denotes the unique C^* algebra generated by the unitary objects $W(f)$. A particular representation of this algebra is given by assigning the vacuum state to the generators $\langle W(f) \rangle_0 = e^{-(1/2)\|f\|_0^2}$, $\|f\|_0^2 = \sum_{n \geq 1} n|f_n|^2$. Starting with the vacuum Hilbert space representation $\mathcal{A}(S^1)_0 = \pi_0(\mathcal{A}(S^1))$, one easily checks that the formula

$$\langle W(f) \rangle_\alpha := e^{i\alpha f_0} \langle W(f) \rangle_0 \tag{7a}$$

$$\pi_\alpha(W(f)) = e^{i\alpha f_0} \pi_0(W(f)) \tag{7b}$$

defines a state with positive energy, that is, one whose GNS representation for $\alpha \neq 0$ is unitarily inequivalent to the vacuum representation. Its incorporation into the vacuum Hilbert space [7b] is part of the DHR formalism. It is convenient to view this change as the result of an application of an automorphism γ_α on the C^* -Weyl algebra $\mathcal{A}(S^1)$ which is implemented by a unitary charge-generating operator Γ_α in a larger (nonseparable) Hilbert space which contains all charge sectors $H_\alpha = \Gamma_\alpha H_0$, $H_0 \equiv H_{\text{vac}} = \mathcal{A}(S^1)\Omega$:

$$\begin{aligned} \langle W(f) \rangle_\alpha &= \langle \gamma_\alpha(W(f)) \rangle_0 \\ \gamma_\alpha(W(f)) &= \Gamma_\alpha W(f) \Gamma_\alpha^* \end{aligned} \tag{8}$$

$\Gamma_\alpha \Omega = \Omega_\alpha$ describes a state with a rotational homogeneous charge distribution; arbitrary charge distributions ρ_α of total charge α that is, $\int (dz/2\pi i)\rho_\alpha = \alpha$ are obtained in the form

$$\psi_{\rho_\alpha}^\zeta = \eta(\rho_\alpha) W(\hat{\rho}_\alpha^\zeta) \Gamma_\alpha \tag{9}$$

where $\eta(\rho_\alpha)$ is a numerical phase factor and the net effect of the Weyl operator is to change the rotational homogeneous charge distribution into ρ_α . The necessary charge-neutral compensating function ρ_α^ζ in the Weyl cocycle $W(\rho_\alpha^\zeta)$ is uniquely determined in terms of ρ_α up to the choice of one point $\zeta \in S^1$ (the determining equation involves the $\ln z$ function which needs the specification of a branch cut (Schroer 2005)). From this formula, one derives the commutation relations $\psi_{\rho_\alpha}^\zeta \psi_{\rho_\beta}^\zeta = e^{\pm i\pi\alpha\beta} \psi_{\rho_\beta}^\zeta \psi_{\rho_\alpha}^\zeta$ for spacelike separations of the ρ supports; hence, these fields are relatively local (bosonic) for $\alpha\beta = 2\mathbb{Z}$. In particular, if only one

type of charge is present, the generating charge is $\alpha_{\text{gen}} = \sqrt{2N}$ and the composite charges are multiples, that is, $\alpha_{\text{gen}}\mathbb{Z}$. This locality condition providing bosonic commutation relations does not yet ensure the ζ -independence. Since the equation which controls the ζ -change turns out to be

$$\psi_{\rho\alpha}^{\zeta_1} \left(\psi_{\rho\alpha}^{\zeta_2} \right)^* = e^{\pm i\pi\alpha\beta} e^{2\pi i Q\alpha} \quad [10]$$

one achieves ζ -independence by restricting the Hilbert space charges to be “dual” to that of the operators, that is,

$$Q = \left\{ \frac{1}{\sqrt{2N}} \mathbb{Z} \right\}$$

The localized $\psi_{\rho\alpha}^{\zeta_1}$ operators acting on the restricted separable Hilbert space H_{res} generate a ζ -independent extended observable algebra $\mathcal{A}_N(S^1)$ (Schroer) and it is not difficult to see that its representation in H_{res} is reducible and that it decomposes into $2N$ charge sectors

$$\left\{ \frac{1}{\sqrt{2N}} n, n = 0, 1, \dots, N-1 \right\}$$

Hence, the process of extension has led to a charge quantization with a finite (“rational”) number of charges relative to the new observable algebra which is neutral in the new charge counting

$$\frac{1}{\alpha_{\text{gen}}} \mathbb{Z} / \alpha_{\text{gen}} \mathbb{Z} = \mathbb{Z} / \alpha_{\text{gen}}^2 = \mathbb{Z}_{2N}$$

The charge-carrying fields in the new setting are also of the above form [9], but now the generating field carries the charge

$$\int \frac{dz}{2\pi i} \rho_{\text{gen}} = Q_{\text{gen}}$$

which is a $(1/2N)$ fraction of the old α_{gen} . Their commutation relations for disjoint charge supports are “braidal” (or better “plektonic” which is more on par with being bosonic/fermionic). (In the abelian case like the present, the terminology “anyonic” enjoys widespread popularity, but in the present context the “any” does not go well with charge quantization.) These objects considered as operators localized on S^1 do depend on the cut ζ , but using an appropriate finite covering of S^1 this dependence is removed (Schroer 2005). So the field algebra $\mathcal{F}_{\mathbb{Z}_{2N}}$ generated by the charge-carrying fields (as opposed to the bosonic observable algebra \mathcal{A}_N) has its unique localization structure on a finite covering of S^1 . An equivalent description which gets rid of ζ consists in dealing with operator-valued sections on S^1 . The

extension $\mathcal{A} \rightarrow \mathcal{A}_N$, which renders the Hilbert space separable and quantizes the charges, seems to be characteristic for abelian current algebra; in all other models which have been constructed up to now the number of sectors is at least denumerable and in the more interesting ones even finite (rational models). An extension is called maximal if there exists no further extension which maintains the bosonic commutation relation. For the case at hand, this would require the presence of another generating field of the same kind as above, which belongs to an integer N' is relatively local to the first one. This is only possible if N is divisible by a square.

In passing, it is interesting to mention a somewhat unexpected relation between the Schwinger model, whose charges are screened, and the Jordan model. Since the Lagrangian formulation of the Schwinger model is a gauge theory, the analog of the four-dimensional “asymptotic freedom” wisdom would suggest the possibility of “charge liberation” in the short-distance limit of this model. This seems to contradict the statement that the intrinsic content of the Schwinger model (QED₂ with massless Fermions) (after removing a classical degree of freedom) is the QFT of a free massive Bose field and such a simple free field is at first sight not expected to contain subtle information about asymptotic charge liberation. (In its original gauge-theoretical form, the Schwinger model has an infinite vacuum degeneracy. The removal of this degeneracy (restoration of the cluster property) with the help of the “ θ -angle formalism” leaves a massive free Bose field (the Schwinger–Higgs mechanism). As expected in $d=1+1$ the model only possesses this phase.) Well, as we have seen above, the massless limit really does have liberated charges and the short-distance limit of the massive free field is the massless model (Schroer).

As a result of the peculiar bosonization/fermionization aspect of the zero-mass limit of the derivative of the massive free field, Jordan’s model is also closely related to the massless Thirring model (and the related Luttinger model for an interacting one-dimensional electron gas) whose massive version is in the class of factorizing models (see later section). (Another structural consequence of this aspect leads to Coleman’s theorem (Schroer 2005) which connects the Mermin–Wagner no-go theorem for two-dimensional spontaneous continuous symmetry breaking with these zero-mass peculiarities.) The Thirring model is a special case in a vast class of “generalized” multicoupling multicomponent Thirring models, that is, models with 4-fermion interactions. Under this name they were studied in the early 1970s (Schroer) with the aim to identify massless subtheories for which the currents form chiral current algebras.

The counterpart of the potential of the conserved Dirac current in the massive Thirring model is the sine-Gordon field, that is, a composite field which in the attractive regime of the Thirring coupling again obeys the so-called sine-Gordon equation of motion. Coleman gave a supportive argument (Schroer 2005) but some fine points about the range of its validity in terms of the coupling strength remained open. (It was noticed that the current potential of the free massive Dirac Fermion ($g=0$) does not obey the sine-Gordon equation (Schroer 2005).) A rigorous confirmation of these facts was recently given in the bootstrap form-factor setting (Schroer 2005). Massive models which have a continuous or discrete internal symmetry have “disorder” fields which implement a “half-space” symmetry on the charge-carrying field (acting as the identity in the other half-axis) and together with the basic pointlike field form composites which have exotic commutation relations (see the last section).

The Conformal Setting, Structural Results

Chiral theories play a special role within the setting of conformal quantum fields. General conformal theories have observable algebras which live on compactified Minkowski space (S^1 in the case of chiral models) and fulfill the Huygens principle, which in an even number of spacetime dimension means that the commutator is only nonvanishing for lightlike separation of the fields. The fact that this classically expected behavior breaks down for nonobservable conformal fields (e.g., the massless Thirring field) was noticed at the beginning of the 1970s and considered paradoxical at that time (“reverberation” in the timelike (Huygens) region). Its resolution around 1974–75 confirmed that such fields are genuine conformal covariant objects but that some fine points about their causality needed to be addressed. The upshot was the proposal of two different but basically equivalent concepts about globally causal fields. They are connected by the following global decomposition formula:

$$\begin{aligned} A(x_{\text{cov}}) &= \sum A_{\alpha,\beta}(x), \quad A_{\alpha,\beta}(x) \\ &= P_\alpha A(x) P_\beta, \quad Z = \sum e^{id_n} P_\alpha \end{aligned} \quad [11]$$

On the left-hand side, the spacetime point of the field is a point on the universal covering of the conformal compactified Minkowski space. These are fields (Lüescher and Mack 1975) (Schroer 2005) which “live” in the sense of quantum (modular) localization on the universal covering spacetime (or on a finite covering, depending on the “rationality”

of the model) and fulfill the global causality condition previously discovered by I Segal (Schroer 2005). They are generally highly reducible with respect to the center of the covering group. The family of fields on the right-hand side, on the other hand, are fields which were introduced (Schroer and Swieca 1974; Schroer *et al.* 1975) with the aim to have objects which live on the projection $x(x_{\text{cov}})$, that is, on the spacetime of the physics laboratory instead of the “hells and heavens” of the covering (Schroer 2005). They are operator-distributional valued sections in the compactification of ordinary Minkowski spacetime. The connection is given by the above decomposition formula into irreducible conformal blocks with respect to the center Z of the noncompact covering group $\widetilde{\text{SO}}(2, n)$ where α, β are labels for the eigenspaces of the generating unitary Z of the abelian center Z . The decomposition [11] is minimal in the sense that in general there generally will be a refinement due to the presence of additional charge superselection rules (and internal group symmetries). The component fields are not Wightman fields since they annihilate the vacuum if the right-hand projection differs from $P_0 = P_{\text{vac}}$.

Note that the Huygens (timelike) region in Minkowski spacetime has a timelike ordering structure $x \prec y$ or $x \succ y$ (earlier or later). In $d=1+1$, the topology allows in addition a spacelike left-right ordering $x \leq y$. In fact, it is precisely the presence of these two orderings in conjunction with the factorization of the vacuum symmetry group $\widetilde{\text{SO}}(2, 2) \simeq \text{PSL}(2R)_l \otimes \text{PSL}(2, R)_r$, in particular $Z = Z_l \otimes Z_r$, which is at the root of a significant simplification. This situation suggested a tensor factorization into chiral components and led to an extremely rich and successful construction program of two-dimensional conformal QFT as a two-step process: the classification of chiral observable algebras on the light ray and the amalgamation of left-right chiral theories to two-dimensional local conformal QFT. The action on the circular coordinates z is through fractional $\text{SU}(1, 1)$ transformations

$$g(z) = \frac{\alpha z + \beta}{\bar{\beta} z + \bar{\alpha}}$$

whereas the covering group acts on the Mack-Lüescher covering coordinates.

The presence of an ordering structure permits the appearance of more general commutation relations for the above $A_{\alpha,\beta}$ component fields namely

$$\begin{aligned} &A_{\alpha,\beta}(x) B_{\beta,\gamma}(y) \\ &= \sum_{\beta'} R_{\beta,\beta'}^{\alpha,\gamma} B_{\alpha,\beta'}(y) A_{\beta',\gamma}(x), \quad x > y \end{aligned} \quad [12]$$

with numerical R -coefficients which, as a result of associativity and relative commutativity with respect to observable fields, have to obey certain structure relations; in this way, Artin braid relations emerge as a new manifestation of the Einstein causality principle for observables in low-dimensional QFT (Rehren and Schroer 1989) (see Schroer 2005). Indeed, the DHR method to interpret charged fields as charge superselection carriers (tied by local representation theory to the bosonic local structure of observable algebras) leads precisely to such a plektonic statistics structure (Fredenhagen *et al.* 1992, Gabbiani and Froehlich 1993) for systems in low spacetime dimension (*see Symmetries in Quantum Field Theory of Lower Spacetime Dimensions*). With an appropriately formulated adjustment to observables fulfilling the Huygens commutativity, this plektonic structure (but now disconnected from particle/field statistics) is also a possible manifestation of causality for the higher-dimensional timelike structure (Schroer 2005).

The only examples known up to the appearance of the seminal BPZ work (Belavin *et al.* 1984) were the abelian current models of the previous section which furnish a rather poor man's illustration of the richness of the decomposition theory. The floodgates of conformal QFT were only opened after the BPZ discovery of "minimal models," which was preceded by the observation (Friedan *et al.* 1984) that the algebra of the stress-energy tensor came with a new representation structure which was not compatible with an underlying internal group symmetry (*see Symmetries in Quantum Field Theory: Algebraic Aspects*).

An important step in the structural study of chiral models was the recognition that the energy-momentum tensor has the commutation structure of a Lie field (Schroer 2005); in the next section, its algebraic structure and its representation theory will be presented.

Chiral Fields and Two-Dimensional Conformal Models

Let us start with a family which generalizes the abelian model of the previous section. Instead of a one-component abelian current we now take n independent copies. The resulting multicomponent Weyl algebra has the previous form except that the current is n -component and the real function space underlying the Weyl algebra consists of functions with values in an n -component real vector space $f \in LV$ with the standard Euclidean inner product denoted by (\cdot, \cdot) . The local extension now leads to

$(\alpha, \beta) \in 2\mathbb{Z}$, that is, an even-integer lattice \mathcal{L} in V , whereas the restricted Hilbert subspace $H_{\mathcal{L}^*}$ which ensures ζ -independence is associated with the dual lattice $L^*: (\lambda_i, \alpha_k) = \delta_{ik}$ which contains \mathcal{L} . The resulting superselection structure (i.e., the Q -spectrum) corresponds to the finite factor group $\mathcal{L}^*/\mathcal{L}$. For self-dual lattices $\mathcal{L}^* = \mathcal{L}$ (which only can occur if $\dim V$ is a multiple of 8), the resulting observable algebra has only the vacuum sector; the most famous case is the Leech lattice Λ_{24} in $\dim V = 24$, also called the "moonshine" model. The observation that the root lattices of the Lie algebras of types A , B , or E (e.g., $\mathfrak{su}(n)$ corresponding to A_{n-1}) also appear among the even-integral lattices suggests that the nonabelian current algebras associated to those Lie algebras can also be implemented. This turns out to be indeed true as far as the level-1 representations are concerned which brings us to the second family: the nonabelian current algebras of level k associated to those Lie algebras; they are characterized by the commutation relation

$$[J_\alpha(z), J_\beta(z')] = i f_{\alpha\beta}^\gamma j_\gamma(z) \delta(z - z') - \frac{1}{2} k g_{\alpha\beta} \delta'(z - z') \quad [13]$$

where $f_{\alpha\beta}^\gamma$ are the structure constants of the underlying Lie algebra, g their Cartan-Killing form, and k , the level of the algebra, must be an integer in order that the current algebra can be globalized to a loop group algebra. The Fourier decomposition of the current leads to the so-called affine Lie algebras, a special family of Kac-Moody algebras. For $k = 1$, these currents can be constructed as bilinears in terms of the multicomponent chiral Dirac field; there exists also the mentioned possibility to obtain them by constructing their maximal Cartan currents within the above abelian setting and representing the remaining nondiagonal currents as certain charge-carrying ("vertex" algebra) operators. Level- k algebras can be constructed from reducing tensor products of k level-1 currents or directly via the representation theory of infinite-dimensional affine Lie algebras. (The global exponentiated algebras (the analogs to the Weyl algebra) are called loop group algebras.) Either way one finds that, for example, the $SU(2)$ current algebra of level k has (together with the vacuum sector) $k + 1$ sectors (inequivalent representations). The different sectors are already distinguished by the structure of their ground states of the conformal Hamiltonian L_0 . Although the computation of higher point correlation functions for $k > 1$, there is no problem in securing the existence of the algebraic nets which define these chiral models as well as their $k + 1$

representation sectors and to identify their generating charge-carrying fields (primary fields) including their R -matrices appearing in their plektonic commutation relations. It is customary to use the notation $SU(2)_k$ for the abstract operator algebras associated with the current generators [13] and we will denote their $k+1$ equivalence classes of representations by $\mathcal{A}_{SU(2)_k, n}, n=0, \dots, k$, whereas representations of current algebras for higher rank groups require a more complicated labeling (in terms of Weyl chambers).

The third family of models are the so-called minimal models which are associated with the Lie-field commutation structure of the chiral stress-energy tensor which results from the chiral decomposition of a conformally covariant two-dimensional stress-energy tensor

$$[T(z), T(z')] = i(T(z) + T(z'))\delta'(z - z') + \frac{ic}{24\pi}\delta'''(z - z') \quad [14]$$

whose Fourier decomposition yields the Witt–Virasoro algebra, that is, a central extension of the Lie algebra of the $\text{Diff}(S^1)$. (The presence of the central term in the context of QFT (the analog of the Schwinger term) was noticed later; however, the terminology Witt–Virasoro algebra in the physics literature came to mean the Lie algebra of diffeomorphisms of the circle including the central extension.) The first two coefficients are determined by the physical role of $T(z)$ as the generating field density for the Lie algebra of the Poincaré group whereas the central extension parameter $c > 0$ (positivity of the two-point function) for the connection with the generation of the Möbius transformations and the undetermined parameter $c > 0$ (the central extension parameter) is easily identified with the strength of the two-point function. Although the structure of the T -correlation functions resembles that of free fields (in the sense that is an algebraically computable unique set of correlation functions once one has specified the two-point function), the realization that c is subject to a discrete quantization if $c < 1$ came as a surprise. As already mentioned, the observation that the superselection sectors (the positive-energy representation structure) of this algebra did not at all follow the logic of a representation theory of an inner symmetry group generated a lot of attention and stimulated a flurry of publications on symmetry concepts beyond groups (quantum groups). A concept of fundamental importance is the DHR theory of localized endomorphisms of operator algebras and

the concept of operator-algebraic inclusions (in particular, inclusions with conditional expectations – V Jones inclusions).

The $SU(2)_k$ current coset construction (Goddard *et al.* 1985) revealed that the proof of existence and the actual construction of the minimal models is related to that of the $SU(2)_k$ current algebras. Constructing a chiral model does not necessarily mean the explicit determination of the n -point Wightman functions of their generating fields (which for most chiral models remains a prohibitively complicated task) but rather a proof of their existence by demonstrating that these models are obtained from free fields by a series of computational complicated but mathematically controlled operator-algebraic steps as reduction of tensor products, formation of orbifolds under group actions, coset constructions, and a special kind of extensions. The generating fields of the models are nontrivial in the sense of not obeying free-field equations (i.e., not being “on-shell”). The cases where one can write down explicit n -point functions of generating fields are very rare; in the case of the minimal family this is limited to the field theory of the Ising model (Schroer 2005).

To show the power of inclusion theory for the determination of the charge content of theory, let us look at a simple illustration in the context of the above multicomponent abelian current algebra. The vacuum representation of the corresponding Weyl algebra is generated from smooth V -valued functions on the circle modulo constant functions (i.e., functions with vanishing total integral) $f \in LV_0$. These functions equipped with the aforementioned complex structure and scalar product yield a Hilbert space. The I -localized subalgebra is generated by the Weyl image of I -supported functions (class functions whose representing functions are constant in the complement I')

$$\begin{aligned} \mathcal{A}(I) &:= \text{alg}\{W(f) | f \in K(I)\} \\ K(I) &= \{f \in LV_0 | f = \text{const. in } I'\} \end{aligned} \quad [15]$$

The one-interval Haag duality $\mathcal{A}(I)' = \mathcal{A}(I')$ (the commutant algebra equals the algebra localized in the complement) is simply a consequence of the fact that the symplectic complement $K(I)'$ in terms of $\text{Im}(f, g)$ consists of real functions in that space which are localized in the complement, that is, $K(I)' = K(I')$. The answer to the same question for a double interval $I = I_1 \cup I_3$ (think of the first and third quadrant on the circle) does not lead to duality but rather to a genuine inclusion

$$\begin{aligned} K((I_1 \cup I_3)') &= K(I_2 \cup I_4) \subset K(I_1 \cup I_3)' \\ K(I_1 \cup I_3) &\subset K((I_1 \cup I_3)')' \end{aligned} \quad [16]$$

The meaning of the left-hand side is clear; these are functions which are constant in $I_1 \cup I_3$ with the same constant in the two intervals whereas the functions on the right-hand side are less restrictive in that the constants can be different. The conversion of real subspaces into von Neumann algebras by the Weyl functor leads to the algebraic inclusion $\mathcal{A}(I_1 \cup I_3) \subset \mathcal{A}((I_1 \cup I_3)')'$. In physical terms, the enlargement results from the fact that within the charge neutral vacuum algebra a charge split with one charge in I_1 and the compensating charge in I_2 for all values of the (unquantized) charge occurs. A more realistic picture is obtained if one allows a charge split to be subjected to a charge quantization implemented by a lattice condition $f(I_2) - f(I_4) \in 2\pi L$ which relates the two multicomponent constant functions (where $f(I)$ denotes the constant value f takes in I). As in the previous one-component case, the choice of even lattices corresponds to the local (bosonic) extensions. Although imposing such a lattice structure destroys the linearity of the K , the functions still define Weyl operators which generated operator algebras $\mathcal{A}_L(I_1 \cup I_2)$. (The linearity structure is recovered on the level of the operator algebra.) But now the inclusion involves the dual lattice L^* (which of course contains the original lattice),

$$\begin{aligned} \mathcal{A}_L(I_1 \cup I_2) &\subset \mathcal{A}_{L^*}(I_1 \cup I_2) \\ \text{ind}\{\mathcal{A}_L(I_1 \cup I_2) \subset \mathcal{A}_L((I_1 \cup I_2)')'\} &= |G| \\ \mathcal{A}_L(I_1 \cup I_2) &= \text{inv}_G \mathcal{A}_{L^*}(I_1 \cup I_2) \end{aligned}$$

This time the possible charge splits correspond to the factor group $G = L^*/L$, that is, the number of possibilities is $|G|$ which measures the relative size of the bigger algebra in terms of the smaller. This is a special case of the general concept of the so-called Jones index of an inclusion which is a numerical measure of its depth. A prerequisite is that the inclusion permits a conditional expectation which is a generalization of the averaging under the “gauge group” G on $\mathcal{A}_{L^*}(I_1 \cup I_2)$ in the third equation above, which identifies the invariant smaller algebra with the fix-point algebra (the invariant part) under the action of G . In fact, using the conceptual framework of Jones, one can show that the two-interval inclusion is independent of the position of the disjoint intervals characterized by the group G .

There exists another form of this inclusion which is more suitable for generalizations. One starts from the charge quantized extended local algebra $\mathcal{A}_L^{\text{ext}} \supset \mathcal{A}$ described earlier in terms of an even-integer lattice L (which lives in the separable Hilbert space H_L) as

our observable algebra. Again the Haag duality is violated and converted into an inclusion $\mathcal{A}_L^{\text{ext}}(I_1 \cup I_2) \subset \mathcal{A}_L^{\text{ext}}((I_1 \cup I_2)')'$ which turns out to have the same $G = L^*/L$ charge structure (it is in fact isomorphic to the previous inclusion). In the general setting (current algebras, minimal model algebras, ...), this double interval inclusion is particularly interesting if the associated Jones index is finite. One finds Kawahigashi *et al.* (2001) (Schroer 2005).

Theorem 1 *A chiral theory with finite Jones index $\mu = \text{ind}\{\mathcal{A}((I_1 \cup I_2)')' : \mathcal{A}(I_1 \cup I_2)\}$ for the double interval inclusion (always assuming that $A(S^1)$ is strongly additive and split) is a rational theory and the statistical dimensions d_ρ of its charge sectors are related to μ through the formula*

$$\mu = \sum_{\rho} d_{\rho}^2 \quad [17]$$

Instead of presenting more constructed chiral models, it may be more informative to mention some of the algebraic methods by which they are constructed and explored. The already mentioned DHR theory provides the conceptual basis for converting the notion of positive-energy representation sectors of the chiral model observable algebras \mathcal{A} (equivalence classes of unitary representations) into localized endomorphisms ρ of this algebra. This is an important step because contrary to group representations which have a natural tensor product composition structure, representations of operator algebras generally do not come with a natural composition structure. The DHR endomorphisms theory of \mathcal{A} leads to fusion laws and an intrinsic notion of generalized statistics (for chiral theories: plektonic in addition to bosonic/fermionic). The chiral statistics parameters are complex numbers (Haag 1992) whose phase is related to a generalized concept of spin via a spin-statistics theorem and whose absolute value (the statistics dimension) generalized the notion of multiplicities of fields known from the description of inner symmetries in higher-dimensional standard QFTs. The different sectors may be united into one bigger algebra called the exchange algebra \mathcal{F}_{red} in the chiral context (the “reduced field bundle” of DHR) in which every sector occurs by definition with multiplicity 1 and the statistics data are encoded into exchange (commutation) relations of charge-carrying operators or generating fields (“exchange algebra fields”) (Schroer 2005). Even though this algebra is useful in that all properties concerning fusion and statistics are nicely encoded, it lacks some cherished properties of standard field theory

namely there is no unique state-field relation, that is, no Reeh-Schlieder property (a field $A_{\alpha\beta}$ whose source projection P_β does not coalesce with the vacuum projection annihilates the vacuum); in operator-algebraic terms, the local algebras are not factors. This poses the question of how to manufacture from the set of all sectors natural (not necessarily local) extensions with these desired properties. It was found that this problem can be characterized in operator-algebraic terms by the existence of the so-called DHR triples (Schroer). In case of rational theories, the number of such extensions is finite and in the aforementioned “classical” current algebra and minimal models they all have been constructed by this method (thus confirming existing results completing the minimal family by adding some missing models). The same method adapted to the chiral tensor product structure of $d=1+1$ conformal observables classifies and constructs all two-dimensional local (bosonic/fermionic) conformal QFT \mathcal{B}_2 which can be associated with the observable chiral input. It turns out that this approach leads to another of those pivotal numerical matrices which encode structural properties of QFT: the coupling matrix Z ,

$$\begin{aligned} \mathcal{A} \otimes \mathcal{A} &\subset \mathcal{B}_2 \\ \sum_{\rho\sigma} Z_{\rho,\sigma} \rho(\mathcal{A}) \otimes \sigma(\mathcal{A}) &\subset \mathcal{A} \otimes \mathcal{A} \end{aligned} \quad [18]$$

where the second line is an inclusion solely expressed in terms of observable algebras from which the desired (isomorphic) inclusion in the first line follows by a canonical construction, the so-called Jones basic construction. The numerical matrix Z is an invariant closely related to the so-called “statistics character matrix” (Schroer 2005) and in case of rational models it is even a modular invariant with respect to the modular $SL(2, \mathbb{Z})$ group transformations (which are closely related to the matrix S in the final section).

Integrability, the Bootstrap Form-Factor Program

Integrability in QFT and the closely associated bootstrap form-factor construction of a very rich class of massive two-dimensional QFTs can be traced back to two observations made during the 1960s and 1970s ideas. On the one hand, there was the time-honored idea to bypass the “off-shell” field-theoretic approach to particle physics in favor of a pure on-shell S -matrix setting which (in particular recommended for strong interactions), as a result of

the elimination of short distances via the mass-shell restriction, would be free of ultraviolet divergencies. This idea was enriched in the 1960s by the crossing property which in turn led to the bootstrap idea, a highly nonlinear seemingly self-consistent proposal for the determination of the S -matrix. However, the protagonists of this S -matrix bootstrap program placed themselves into a totally antagonistic fruitless position with respect to QFT so that the strong return of QFT in the form of gauge theory undermined their credibility. On the other hand, there were rather convincing quasiclassical calculations in certain two-dimensional massive QFTs as, for example, the sine-Gordon model which indicated that the obtained quasiclassical mass spectrum is exact and hence suggested that the associated QFTs are integrable (Dashen *et al.* 1975) and have no real particle creation. These provocative observations asked for a structural explanation beyond quasiclassical approximations, and it soon became clear that the natural setting for obtaining such mass formulas was that of the “fusion” of boundstate poles of unitary crossing-symmetric purely elastic S -matrices; first in the special context of the sine-Gordon model (Schroer *et al.* 1976) and later as a classification program from which factorizing S -matrices can be determined by solving well-defined equations for the elastic two-particle S -matrix (Karowski *et al.* 1977). (It was incorrectly believed that the “nontrivial elastic scattering implies particle creation” statement of Aks (Aks, 1963) is also valid for low-dimensional QFTs.) Some equations in this bootstrap approach resembled mathematical structures which appeared in C N Yang’s work on nonrelativistic δ -function particle interactions as well as relations for Boltzmann weights in Baxter’s work on solvable lattice models; hence, they were referred to as Yang-Baxter relations. These results suggested that the old bootstrap idea, once liberated from its ideological dead freight (in particular from the claim that the bootstrap leads to a unique “theory of everything” (minus gravity)), generates a useful setting for the classification and construction of factorizing two-dimensional relativistic S -matrices. Adapting certain known relations between two-particle form factors of field operators and the S -matrix to the case at hand (Karowski and Weisz 1978), and extending this with hindsight to generalized (multiparticle) form factors, one arrived at the axiomatized recipes of the bootstrap form-factor program of $d=1+1$ factorizable models (Smirnov 1992). Although this approach can be formulated within the

setting of the LSZ scattering formalism, the use of a certain algebraic structure (Zamolodchikov and Zamolodchikov 1979) which in the simplest version reads

$$\begin{aligned} Z(\theta)Z^*(\theta') &= S^{(2)}(\theta - \theta')Z^*(\theta')Z(\theta) + \delta(\theta - \theta') \\ Z(\theta)Z(\theta') &= S^{(2)}(\theta' - \theta)Z(\theta')Z(\theta) \end{aligned} \tag{19}$$

(the δ -term Faddeev is due to Faddeev) brought significant simplifications. In the general case, the Z 's are vector valued and the $S^{(2)}$ -structure function is matrix valued. (The identification of the Z -F structure coefficients with the elastic two-particle S -matrix $S^{(2)}$ (which is preemtped by our notation) can be shown to follow from the physical interpretation of the Z -F structure in terms of localization.) In that case the associativity of the Z -F algebra is equivalent to the Yang-Baxter equations. Recently, it became clear that this algebraic relation has a deep physical interpretation; it is the simplest algebraic structure which can be associated with generators of nontrivial wedge-localized operator algebras (see the next section).

Conceptually as well as computationally it is much simpler to identify the intrinsic meaning of integrability in QFT with the factorization of its S -matrix or a certain property of wedge-localized algebras (see next section) than to establish integrability (see Integrability and Quantum Field Theory).

The first step of the bootstrap form-factor program namely the classification and construction of model S -matrices follows a combination of two patterns: prescribing particle multiplets transforming according to group symmetries and/or specifying structural properties of the particle spectrum. The simplest illustration for the latter strategy is supplied by the \mathbb{Z}_N model. In terms of particle content, \mathbb{Z}_N demands the identification of the N th bound state with the antiparticle. Since the fusion condition for the bound mass $m_b^2 = (p_1 + p_2)^2 = m_1^2 + m_2^2 + 2m_1 m_2 ch(\theta_1 - \theta_2)$ is only possible for a pure imaginary rapidity difference $\theta_{12} = \theta_1 - \theta_2 = i\alpha$ ("binding angle"). Hence, the binding of two "elementary" particles of mass m gives

$$m_2 = m \frac{\sin 2\alpha}{\sin \alpha}$$

and more generally of k particles with

$$m_k = m \frac{\sin k\alpha}{\sin \alpha}$$

so that the antiparticle mass condition $m_N = \bar{m} = m$ fixes the binding angle to $\alpha = 2\pi/N$. (The quotation mark is meant to indicate that in contrast to the Schrödinger QM there is "nuclear democracy" on

the level of particles. The inexorable presence of interaction-caused vacuum polarization limits a fundamental/fused hierarchy to the fusion of charges.) The minimal (no additional physical poles) two-particle S -matrix in terms of which the n -particle S -matrix factorizes is therefore

$$S_{\min}^{(2)} = \frac{\sin(1/2)(\theta + (2\pi i)/N)}{\sin(1/2)(\theta - (2\pi i)/N)} \tag{20}$$

(minimal=without so-called CDD poles) The $SU(N)$ model as compared with the $U(N)$ model requires a similar identification of bound states of $N - 1$ particles with an antiparticle. This S -matrix enters as in the equation for the vacuum to n -particle meromorphic form factor of local operators; together with the crossing and the so-called "kinematical pole equation," one obtains a recursive infinite system linking a certain residue with a form factor involving a lower number of particles. The solutions of this infinite system form a linear space from which the form factors of specific tensor fields can be selected by a process which is analogous but more involved than the specification of a Wick basis of composite free fields. Although the statistics property of two-dimensional massive fields is not intrinsic but a matter of choice, it would be natural to realize, for example, the \mathbb{Z}_N fields as \mathbb{Z}_N -anyons.

Another rich class of factorizing models are the Toda theories of which the sine-Gordon and sinh-Gordon are the simplest cases. For their descriptions, the quasiclassical use of Lagrangians (supported by integrability) turns out to be of some help in setting up their more involved bootstrap form-factor construction.

The unexpected appearance of objects with new fundamental (solitonic) charges (e.g., the Thirring field as the carrier of a solitonic sine-Gordon charge) and the unexpected confinement of charges (e.g., the $CP(1)$ model as a confined $SU(2)$ model) turn out to be opposite sides of the same coin and both cases have realizations in the setting of factorizing models (Schroer 2005).

Recent Developments

There are two ongoing developments which place the two-dimensional bootstrap form-factor program into a more general setting which permits to understand its position in the general context of local quantum physics.

One of these starts from the observation that the smallest spacetime localization region in which it is possible to find vacuum-polarization-free generators (PFG) in the presence of interactions is the wedge

region. If one demands in addition that these generators (necessarily unbounded operators) have the standard domain properties of QFT (which include stability of the domain under translations), then one finds that this leads precisely to the two-dimensional Z-F algebraic structure which in turn in this way a spacetime interpretation for the first time acquires. In these investigations (Schroer 2005), modular localization theory plays a prominent role and there are strong indications that with these methods one can show the nontriviality of intersections of wedge algebras which is the algebraic criterion for the existence of a model within local quantum physics.

There is a second constructive idea based on light-front holography which uses the radical reorganization of spacetime properties of the algebraic structure while maintaining the physical content including the Hilbert space. Since spacetime localization aspects (apart from the remark about wedge algebras and their PFG generators made before) are traditionally related to the concept of fields, holographic methods tend to de-emphasize the particle structure in favor of “field properties.” Indeed, the transversely extended chiral theories which arise as the holographic image lead to simplification of many interesting properties with very similar aims to the old “light-cone quantization” except that light-front holography is another way of looking at the original local ambient theory without subjecting it to another quantization. (The price for this simplification is that as a result of the nonuniqueness of the holographic inversion certain problems cannot be formulated.)

Actually, as a result of the absence of a transverse direction in the two-dimensional setting, the family of factorizing models provides an excellent theoretical laboratory to study their rigorous “chiral encoding” which is conceptually very different from Zamolodchikov’s perturbative relation (which is based on identifying a factorizing model in terms of a perturbation on a chiral theory).

It turns out that the issue of statistics of particles loses its physical relevance for two-dimensional massive models since they can be changed without affecting the physical content. Instead such notions as order/disorder fields and soliton take their place (Schroer 2005).

In accordance with its historical origin, the theory of two-dimensional factorizing models may also be viewed as an outgrowth of the quantization of classical integrable systems (Integrability and Quantum Field Theory). But in comparison with the rather involved structure of integrability (verifying the existence of sufficiently many commuting conservation laws), the conceptual setting of factorizing

models within the scattering framework (factorization follows from existence of wedge-localized tempered PFGs) is rather simple and intrinsic (Schroer 2005).

Among the additional ongoing investigations in which the conceptual relation with higher-dimensional QFT is achieved via modular localization theory, we will select three which have caught our, active attention. One is motivated by the recent discovery of the adaptation of Einstein’s classical principle of local covariance to QFT in curved spacetime. The central question raised by this work (*see Algebraic Approach to Quantum Field Theory*) is if all models of Minkowski spacetime QFTs permit a local covariant extension to curved spacetime and if not which models do? In the realm of chiral QFT, this would amount to ask if all Moebius-invariant models are also $\text{Diff}(S^1)$ -covariant. It has been known for sometime that a QFT with all its rich physical content can be uniquely defined in terms of a carefully chosen relative position of a finite number of copies of one unique von Neumann operator algebra within one common Hilbert space. This is a perfect quantum field-theoretical illustration for Leibniz’s philosophical proposal that reality results from the relative position of “monades” (As opposed to the more common (Newtonian) view that the material reality originates from a material content being placed into a spacetime vessel) if one takes the step of identifying the hyperfinite type III_1 Murray von Neumann factor algebra with an abstract monade from which the different copies result from different ways of positioning in a shared Hilbert space (Schroer 2005). In particular, Moebius-covariant chiral QFTs arise from two monades with a joint intersection defining a third monade in such a way that the relative positions are specified in terms of natural modular concepts (without reference to geometry). This begs the question whether one can extend these modular-based algebraic ideas to pass from the global vacuum preserving Moebius invariance to local $\text{Diff}(S)$ covariance $\text{Moeb} \rightarrow \text{Diff}(S^1)$. This would be precisely the two-dimensional adaptation of the crucial problem raised by the recent successful generalization of the local covariance principle underlying Einstein’s classical theory of gravity to QFT in curved spacetime: does every Poincaré covariant Minkowski spacetime QFT allow a unique correspondence with one curved spacetime (having the same abstract algebraic substrate but with a totally different spacetime encoding)? In the chiral context, one is led to the notion of “partially geometric modular groups” which only act geometrically if restricted to specific subalgebras (Schroer

2005). It is hard to imagine how one can combine quantum theory and gravity without understanding first the still mysterious links between spacetime geometry, thermal properties, and relative positioning of monades in a joint Hilbert space.

A second important umbilical cord with higher-dimensional theories is the issue of “Euclideanization” in particular the chiral counterpart of Osterwalder–Schrader localization and the closely related Nelson–Symanzik duality. In concrete chiral models (e.g., the models in the section “Chiral fields and two-dimensional conformal models”), it has been noted as a result of explicit calculations that the analytic continuation in the angular parametrization for thermal correlation functions leads to a duality relation in

$$\begin{aligned} \langle A(\varphi_1, \dots, \varphi_n) \rangle_{\alpha, 2\pi\beta_t} \\ = \left(\frac{i}{\beta_t} \right)^a \sum_{\gamma} S_{\alpha\gamma} \left\langle A \left(\frac{i}{\beta_t} \varphi_1, \dots, \frac{i}{\beta_t} \varphi_n \right) \right\rangle_{\gamma, (2\pi/\beta_t)} \end{aligned} \quad [21]$$

where the thermal correlation function is defined as

$$\begin{aligned} \langle A(\varphi_1, \dots, \varphi_n) \rangle_{\rho_\alpha, 2\pi\beta_t} \\ := \text{tr}_{H_{\rho_\alpha}} e^{-2\pi\beta_t(L_0^{\rho_\alpha} - (c/24))} \pi_{\rho_\alpha}(A(\varphi_1, \dots, \varphi_n)) \end{aligned} \quad [22]$$

$$A(\varphi_1, \dots, \varphi_n) = \prod_{i=1}^n A_i(\varphi_i)$$

Compared with the thermally extended Nelson–Symanzik relation for two-dimensional QFT one notices that in addition to the expected behavior of real coordinates becoming imaginary and the 2π -periodicity changing role with the (suitably normalized) KMS inverse temperature, there is a rotation in the space of superselected charges in terms of a unitary matrix S whose origin lies in the braid group statistics (the statistics character matrix). The deeper structural explanation which shows that this relation is not just a property of special models, but rather a generic property of chiral QFT, comes from a very deep angular Euclideanization which is based on modular theory (Schroer). Specializing $A = \text{identity}$, one obtains a relation for the partition function, the famous Verlinde identity which is part of the transformation law of the thermal angular correlation functions under the $SL(2, \mathbb{R})$ modular group.

There are many additional important observations on factorizing models whose relation to the physical principles of QFT, unlike the bootstrap form-factor program, is not yet settled. The meaning of the c -parameter outside the chiral setting and ideas on its renormalization group flow as well as the various formulations of the thermodynamic Bethe ansatz

belong to a series of interesting observations whose final relation to the principles of QFT still needs clarification.

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Bosons and Fermions in External Fields; Euclidean Field Theory; Integrability and Quantum Field Theory; Operator Product Expansion in Quantum Field Theory; Sine-Gordon Equation; Symmetries in Quantum Field Theory: Algebraic Aspects; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Tomita–Takesaki Modular Theory.

Further Reading

- Abdalla E, Abdalla MCB, and Rothe K (1991) *Non-Perturbative Methods in 2-Dimensional Quantum Field Theory*. Singapore: World Scientific.
- Belavin AA, Polyakov AM, and Zamolodchikov AB (1984) Infinite conformal symmetry in two-dimensional quantum field theory. *Nuclear Physics B* 241: 333.
- Bisognano JJ and Wichmann EH (1975) *Journal of Mathematical Physics* 16: 985.
- Dashen F, Hasslacher B, and Neveu A (1975) *Physics Reviews D* 11: 3424.
- Di Francesco P, Mathieu P, and Sénéchal D (1996) *Conformal Field Theory*. Berlin: Springer.
- Doplicher S, Haag R, and Roberts JE (1971/1974) *Communications in Mathematical Physics* 23: 199; 35: 49.
- Fredenhagen K, Rehren KH, and Schroer B (1992) Superselection sector with braid group statistics and exchange algebras II: Geometric aspects and conformal invariance. *Reviews of Mathematical Physics* 1 (special issue): 113.
- Furlan P, Sotkov G, and Todorov I (1989) Two-dimensional conformal quantum field theory. *Rivista del Nuovo Cimento* 12: 1–203.
- Gabbiani F and Froehlich J (1993) Operator algebras and conformal field theory. *Communications in Mathematical Physics* 155: 569.
- Glimm J and Jaffe A (1987) *Quantum Physics. A Functional Integral Point of View*. Berlin: Springer.
- Ginsparg P (1990) Applied conformal field theory. In: Brezin E and Zinn-Justin J (eds.) *Fields, Strings and Critical Phenomena*, Les Houches 1988. Amsterdam: North-Holland.
- Goddard P, Kent A, and Olive D (1985) Virasoro algebras and coset space models. *Physics Letters B* 152: 88.
- Haag R (1992) *Local Quantum Physics*. Berlin: Springer.
- Ising E (1925) *Zeitschrift für Physik* 31: 253.
- Jordan P (1937) Beiträge zur Neutrintheorie des Lichts. *Zeitschrift für Physik* 114: 229 and earlier papers quoted therein.
- Karowski M, Thun H-J, Truong TT, and Weisz P (1977) *Physics Letters B* 67: 321.
- Karowski M and Weisz P (1978) *Physics Reviews B* 139: 445.
- Kawahigashi Y, Longo R, and Mueger M (2001) Multi-interval subfactors and modularity in representations of conformal field theory. *Communications in Mathematical Physics* 219: 631.
- Lenz W (1920) *Physikalische Zeitschrift* 21: 613.
- Lüescher M and Mack G (1975) Global conformal invariance in quantum field theory. *Communications in Mathematical Physics* 41: 203.
- Rehren K-H and Schroer B (1987) Exchange algebra and Ising n -point functions. *Physics Letters B* 198: 84.

- Rehren KH and Schroer B (1989) Einstein causality and Artin braids. *Nuclear Physics B* 312: 715.
- Schroer B (2005) Two-dimensional models, a testing ground for principles and concepts of QFT, *Annals of Physics* (in print) (hep-th/0504206).
- Schroer B and Swieca JA (1974) Conformational transformations for quantized fields. *Physics Reviews D* 10: 480.
- Schroer B, Swieca JA, and Voelkel AH (1975) Global operator expansions in conformally invariant relativistic quantum field theory. *Physics Reviews D* 11: 11.
- Schroer B, Truong TT, and Weisz P (1976) Towards an explicit construction of the Sine-Gordon field theory. *Annals of Physics (New York)* 102: 156.
- Schwinger J (1962) *Physical Review* 128: 2425.
- Schwinger J (1963) Gauge theory of vector particles. In: *Theoretical Physics Trieste Lectures 1962*. Wien: IAEA.
- Smirnov FA (1992) *Advanced Series in Mathematical Physics* 14. Singapore: World Scientific.
- Streater RF and Wightman AS (1964) *PCT, Spin and Statistics and All That*. New York: Benjamin.
- Zamolodchikov AB and Zamolodchikov AB (1979) *Annals of Physics (New York)* 120: 253.

U

Universality and Renormalization

M Lyubich, University of Toronto, Toronto, ON, Canada and Stony Brook University, Stony Brook, NY, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Discovery of the universality phenomenon and the underlying renormalization mechanism by Feigenbaum and independently by Coullet and Tresser in late 1970s was one of the most influential events in the dynamical systems theory in the last quarter of the twentieth century. It was numerically observed that the cascades of doubling bifurcations leading to chaotic regimes in one-parameter families of interval maps, as well as the dynamical attractors that appear in the limits, exhibit the universal small-scale geometry. To explain this surprising observation, a “Renormalization Conjecture” was formulated which asserted that a natural renormalization operator acting in the space of dynamical systems has a unique hyperbolic fixed point.

It took about two decades to prove this conjecture rigorously (and without the help of computers). The proof revealed rich mathematical structures behind the universality phenomenon that linked it tightly to holomorphic dynamics and conformal and hyperbolic geometry.

Besides the universality *per se*, the renormalization theory led to many other important results. It includes the proof of the regular or stochastic dichotomy that gives us a complete understanding of the real quadratic family (and more general families of one-dimensional maps) from measure-theoretic point of view, as well as deep advances in several key problems of holomorphic dynamics.

Since the original discovery, many other manifestations of the universality have been observed, experimentally, numerically, and theoretically, in various classes of dynamical systems. However, in this article we will focus on mathematical aspects of the original phenomenon.

General Terminology and Notations

We will use general notations and terminology from Holomorphic Dynamics.

Unimodal Maps

Definitions and Conventions

Let us consider a smooth interval map $f: I \rightarrow I$. It is called *unimodal* if it has a single critical point c and this point is an extremum. We assume that the critical point is nondegenerate, unless otherwise it is explicitly stated. A unimodal map is called *S-unimodal* if it has a negative Schwarzian derivative:

$$Sf = \frac{f'''}{f'} - \frac{3}{2} \left(\frac{f''}{f'} \right)^2 < 0$$

For simplicity, we also assume that the map f is even, and normalize it so that $c=0$ and one of the endpoints of I is a fixed point.

Topological Dynamics

Let $J \ni 0$ be a 0-symmetric periodic interval, that is, $f^p(J) \subset J$ for some $p \in \mathbb{N}$, such that the intervals $J_k = f^k(J)$, $k=0, 1, \dots, p-1$, have disjoint interiors. Then we refer to $\cup J_k$ as a *cycle of intervals* of period p .

According to their topological dynamics, *S-unimodal* maps can be divided into three possible types (Sharkovskii, Singer, Guckenheimer, Misiurawicz, van Strien, Blokh, etc.):

- *Regular maps*. Such a map has an attracting or parabolic cycle α . In this case, almost all trajectories of f converge to α . In case α is attracting, the map f is also called *hyperbolic* (see Holomorphic Dynamics).
- *Topologically chaotic maps*. For such a map, there is a cycle of intervals $\cup J_k$ such that the restriction $f|_{\cup J_k}$ is *topologically transitive* (i.e., it has a dense orbit). Moreover, for almost all $z \in I$, $\text{orb } z$ eventually lands in this cycle.
- *Infinitely renormalizable maps*. For such a map, there is a nested sequence of periodic intervals $J^1 \supset J^2 \supset \dots \ni 0$ of periods $p_n \rightarrow \infty$. Then the

intersection of the corresponding cycles of intervals,

$$A = A_f = \bigcap_{n=0}^\infty \bigcup_{k=0}^{p_n-1} f^k(J^n) \tag{1}$$

is a Cantor set endowed with a natural group structure (inverse limit of cyclic groups $\mathbb{Z}/p_n\mathbb{Z}$) such that $f|_A$ becomes a group translation. Moreover, $f^n z \rightarrow A$ for a.e. $z \in I$. This Cantor set is also called the *Feigenbaum attractor* of f .

Kneading Theory

Kneading theory (Milnor and Thurston, mid-1970s) gives a complete topological classification of S -unimodal maps (and more general one-dimensional maps). Let I_+ and I_- stand for the components of $I \setminus \{0\}$, where $I_+ \ni f(0)$. To any point $x \in I$, let us associate its *itinerary* $(\varepsilon_n)_{n=0}^N$, where $\varepsilon_n \in \{+, -, 0\}$, $N \in \mathbb{Z}_+ \cup \infty$, in the following way. If x is precritical then $N \in \mathbb{Z}_+$ is the smallest number such that $f^N x = 0$, and we let $\varepsilon_N = 0$. Otherwise, $N = \infty$. For $n < N$, $\varepsilon_n = +$ if $f^n x \in I_+$, and $\varepsilon_n = -$ if $f^n x \in I_-$.

The *kneading sequence* of f is the itinerary of the critical value $f(0)$. It essentially classifies S -unimodal maps: *two nonregular S -unimodal maps are topologically conjugate if and only if they have the same kneading sequence.* (In the regular case, one should state if the map is hyperbolic or parabolic and specify the sign of the multiplier of the corresponding cycle.)

The kneading theory completely describes *admissible* kneading sequences (realizable by some unimodal maps), and order them linearly in such a way that a bigger sequence corresponds to a more “complicated” map. The minimal admissible kneading sequence, $+++$, is realized by the parabolic map $x \mapsto x^2 + 1/4$, while the maximal one, $+-+--$, is realized by the Chebyshev map $x \mapsto x^2 - 2$.

A central result of the kneading theory is the *Intermediate Value Theorem* asserting that a smooth one-parameter family of S -unimodal maps f_t containing two kneading sequences also contains all intermediate kneading sequences. In particular, a family that contains the above maximal and the minimal kneading sequences, contains *all* admissible kneading sequences. Such a family is called *full*. We see that the real quadratic family $P_c, c \in [-2, 1/4]$, is full: *any S -unimodal map is topologically equivalent to some quadratic polynomial.* This indicates dynamical significance of the quadratic family.

We say that a one-parameter family of unimodal maps f_t is *almost full* if it contains all admissible kneading sequences except possibly the minimal one.

Universality Phenomenon

Universal Geometry of Doubling Bifurcations and the Feigenbaum Attractor

Let us consider the real quadratic family $P_c : x \mapsto x^2 + c$, $c \in [-2, 1/4]$. As the parameter c moves down from $1/4$, we observe a sequence of doubling bifurcations c_n where the attracting cycle of period 2^n gives birth to an attracting cycle of period 2^{n+1} , $n = 0, 1, \dots$ (see Holomorphic Dynamics and Figure 1). This sequence converges to the *Feigenbaum parameter* c_∞ at exponential rate: $c_n - c_\infty \sim \lambda^{-n}$, where $\lambda \approx 4.6$. It turns out that if we consider a similar one-parameter family of unimodal maps, say $x \mapsto a \sin x$, we observe a similar sequence of doubling bifurcations converging to the limit exponentially at the *same* rate λ^{-n} , independently of the family under consideration.

In the dynamical space, let us consider the Feigenbaum attractor A_f [1] of an infinitely renormalizable S -unimodal map f that appear in the limit of doubling bifurcations (so that the periods of periodic intervals J^n are equal to 2^n). Let us consider the *scaling factors* $\sigma_n = |J^n|/|J^{n-1}|$. Then $\sigma_n \rightarrow \sigma_\infty$, where the limiting scaling factor $\sigma_\infty \approx 2.6$ is

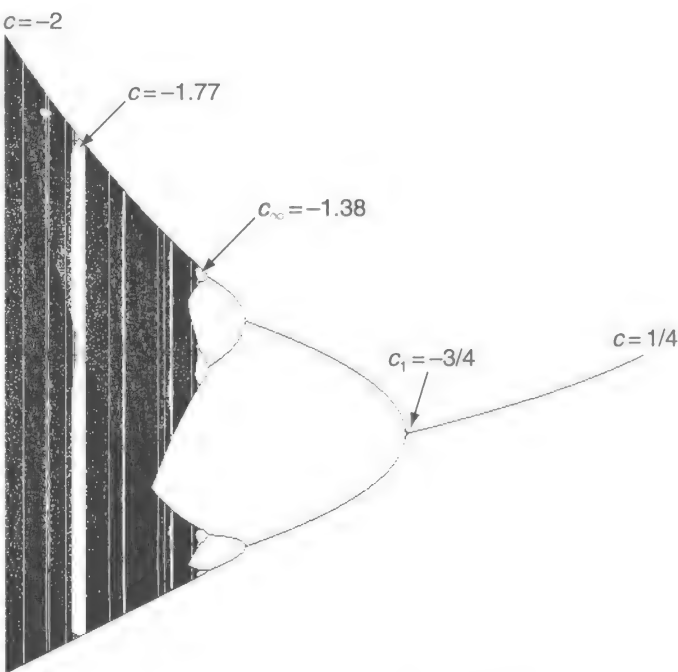


Figure 1 Real quadratic family $P_c : x \mapsto x^2 + c$. This picture presents how the limit set of the orbit $\{P_c^n(0)\}_{n=0}^\infty$ bifurcates as the parameter c changes from $1/4$ on the right to -2 on the left. Three topological types of regimes are intertwined in an intricate way. The gaps correspond to the regular regimes. The black regions correspond to the chaotic regimes (though, of course, there are many narrow invisible gaps therein). In the beginning (on the right) one can see the cascade of doubling bifurcations. This picture became symbolic for one-dimensional dynamics.

independent of the particular map f under consideration. Thus, the small-scale geometry of A_f is universal.

This was historically the first observed manifestation of the *quantitative universality* of dynamical and parameter structures.

Feigenbaum–Coullet–Tresser Renormalization Conjecture

To explain the above universality phenomenon, Feigenbaum and independently Coullet and Tresser, formulated the following Renormalization Conjecture. Let us consider the space \mathcal{U} of S -unimodal maps $f : [-1, 1] \rightarrow [-1, 1]$. A map $f \in \mathcal{U}$ is called (doubling) *renormalizable* if it has a cycle of intervals $J \rightarrow J_1 \rightarrow J$ of period 2. Then, for any $n \in \mathbb{Z}_+ \cup \{\infty\}$, we can naturally define n -times renormalizable maps, where $n = 0$ corresponds to the *non-renormalizable* case, while $n = \infty$ corresponds to the *infinitely renormalizable* case.

Let $\mathcal{U}' \subset \mathcal{U}$ be the space of doubling renormalizable maps. If $f \in \mathcal{U}'$ then $f^2 : J \rightarrow J$ is an S -unimodal map as well, and we define the (doubling) *renormalization operator* $R : \mathcal{U}' \rightarrow \mathcal{U}$ as the rescaling of this map:

$$Rf(x) = \sigma^{-1}f^2(\sigma x)$$

where $\sigma = |J|/2$.

The Renormalization Conjecture asserted that:

- The renormalization operator R has a unique fixed point f_* , and this point is hyperbolic;
- the stable manifold $W^s(f_*)$ consists of infinitely renormalizable unimodal maps;
- the unstable manifold $W^u(f_*)$ is one dimensional and represents an almost full family of unimodal maps (see the section “Kneading theory”); and
- the quadratic family $\{P_c\}$ transversally intersects $W^s(f_*)$ (see Figure 2).

Assuming this conjecture, one can see that for any curve $t \mapsto g_t$ in \mathcal{U} that transversally intersects the stable manifold $W^s(f_*)$ at some moment t_* , the doubling bifurcations parameters t_n converge to t_* at

exponential rate λ^{-n} , where λ is the unstable eigenvalue of the differential $DR(f_*)$. This explains the universal geometry of doubling bifurcations.

One can also show that the Feigenbaum attractor A_f of any map $f \in W^s(f_*)$ is *smoothly equivalent* to A_{f_*} , which explains the universal small-scale geometry of these attractors.

Full Renormalization Horseshoe

Along with period doublings, one can consider period triplings, quadruplings, etc. A unimodal map $f \in \mathcal{U}$ is said to be *renormalizable with period p* if it has a cycle of intervals $J \rightarrow J_1 \rightarrow \dots \rightarrow J_{p-1} \rightarrow J$ of period p . The corresponding renormalization operator is defined as $Rf(x) = \sigma^{-1}f^p(\sigma x)$, where $\sigma = |J|/2$.

The *combinatorics* or *type τ* of the renormalization operator is the order of the intervals $J_k, k = 0, 1, \dots, p - 1$, on the real line (up to reversal). (For instance, there are three admissible combinatorics τ of period 5.) If we want to specify combinatorics of the renormalization operator under consideration, we use notation R_τ . This operator is defined on the “renormalization strip” \mathcal{U}^τ of unimodal maps $f \in \mathcal{U}$ that are renormalizable with combinatorics τ .

The Renormalization Conjecture admits a straightforward generalization to any renormalization operator R_τ . More interestingly, one can formulate a stronger version of it by putting all the admissible renormalization types together. Let \mathcal{T} stand for the set of all *minimal* renormalization types, that is, the types that cannot be factored through other types. Then the renormalization strips $\mathcal{U}^\tau, \tau \in \mathcal{T}$, are pairwise disjoint, and we can define the *full renormalization operator*

$$R : \bigcup_{\tau \in \mathcal{T}} \mathcal{U}^\tau \rightarrow \mathcal{U} \tag{2}$$

by letting $R|_{\mathcal{U}^\tau} = R_\tau$. Then the strong version of the renormalization conjecture asserted that:

- there is an R -invariant hyperbolic subset $\mathcal{A} \subset \mathcal{U}$ called the *full renormalization horseshoe* such that the restriction $R|_{\mathcal{A}}$ is topologically conjugate to the full shift σ on the space Σ of bi-infinite sequences $(\dots, \tau_{-1}, \tau_0, \tau_1, \dots)$ of symbols $\tau_n \in \mathcal{T}$;
- for any $f_* \in \mathcal{A}$, the stable manifold $W^s(f_*)$ consists of infinitely renormalizable maps $f \in \mathcal{U}$ with the same combinatorics as f_* ;
- for any $f_* \in \mathcal{A}$, the unstable manifold $W^u(f_*)$ is one-dimensional and represents an almost full family of unimodal maps; and
- the real quadratic family $\{P_c\}$ transversally intersects all stable manifolds $W^s(f_*)$.

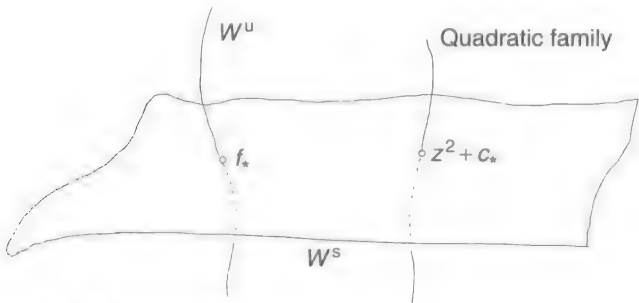


Figure 2 Renormalization fixed point.

Complex Renormalization

Polynomial-Like Maps

A *polynomial-like map* is a holomorphic branched covering of finite degree $f:U\rightarrow U'$, where $U\Subset U'\subset\mathbb{C}$ are topological disks (In other words, the maps f is *proper*, that is, full preimages $f^{-1}(K)$ of compact sets $K\subset U'$ are compact). For instance, if f is a polynomial of degree d then for a sufficiently large radius $R>0$, the map $f:f^{-1}(\mathbb{D}_R)\rightarrow\mathbb{D}_R$ is a polynomial-like map of the same degree d . We refer to such polynomial-like maps as “polynomials.”

The *filled Julia set* of f is the set of nonescaping points:

$$K(f)=\{z:f^n z\in U, n=0,1,\dots\}$$

The Julia set of f is the boundary of its filled Julia set: $J(f)=\partial K(f)$.

A polynomial-like map of degree d has $d-1$ critical points counted with multiplicities. The Julia set (and the filled Julia set) is connected if and only if all the critical points c_i are nonescaping, that is, $c_i\in K(f)$.

A polynomial-like map of degree 2 is called *quadratic-like*. The Julia set of a quadratic-like map is either connected or a Cantor set, depending on whether its critical point is nonescaping or otherwise.

The domain of a polynomial-like map is allowed to be slightly adjusted by taking V' to be a topological disk such that $U\subset V'\subset U'$ and letting $V=f^{-1}(V')$. We say that two polynomial-like maps represent the same germ if one can be obtained from the other by a sequence of such adjustments.

We will be mostly interested in the quadratic case; so let \mathcal{Q} be the space of quadratic-like germs considered up to affine conjugacy, and let \mathcal{C} be the *connectedness locus* in \mathcal{Q} , that is, the subset of $f\in\mathcal{Q}$ with connected Julia set. The space \mathcal{Q} has a natural complex analytic structure such that holomorphic curves in \mathcal{Q} are represented by holomorphic families $f_\lambda(z)$ of quadratic-like maps.

Two polynomial-like maps are called *hybrid equivalent* if they are conjugate by a quasiconformal map h such that $\bar{\partial}h=0$ a.e. on $K(f)$ (in particular, h is conformal on $\text{int } K(f)$). By the *Straightening Theorem*, any polynomial-like map is hybrid equivalent (after an adjustment of its domain) to a polynomial of the same degree (called the “straightening” of f). The straightening depends only on the germ of f .

For a quadratic-like map f with connected Julia set, the straightening $P_c:z\mapsto z^2+c$ is *unique*, $c=\chi(f)$. Thus, we obtain the straightening map

$\chi:\mathcal{C}\rightarrow M$, where M is the Mandelbrot set (see Holomorphic Dynamics). We let $\mathcal{H}_c=\chi^{-1}(c)$ be the hybrid class passing through a point $c\in M$. One can show that \mathcal{H}_c is a *codimension-one submanifold* in \mathcal{Q} .

Any quadratic-like map has two fixed points counted with multiplicity. In the case of connected Julia set, these fixed points have a different dynamical meaning: one of them, called α , is either attracting, or neutral, or repelling *separating*, that is, $J(f)\setminus\{\alpha\}$ is disconnected. Another one, called β , is either parabolic with multiplier 1 (and then it coincides with α) or repelling *nonseparating*.

In what follows, we normalize quadratic-like maps so that 0 is their critical point.

Complex Renormalization and Little Mandelbrot Sets

A quadratic-like map $f:U\rightarrow U'$ with connected Julia set is called *renormalizable* if there is a topological disk $V\ni 0$ and a natural number $p\geq 2$ called the *renormalization period* such that:

- letting $g=f^p|_V$ and $V'=g(V)$, the map $g:V\rightarrow V'$ is quadratic-like;
- the *little Julia set* $K(g)$ is connected; and
- the sets $g^n(K(g)), n=1,\dots,p-1$, can intersect $K(g)$ only at the β -fixed point of g .

Under these circumstances, the quadratic-like germ g considered up to affine conjugacy is called the *renormalization* of the quadratic-like germ f ; $g=Rf$. Moreover, one says that f is *primitively renormalizable* if the little Julia sets $g^n(K(g)), n=1,\dots,p-1$, are pairwise disjoint. Otherwise, f is *satellite renormalizable*.

As in the unimodal case, one can define *combinatorics* or *type* τ of the complex renormalization. Roughly speaking, renormalizable maps with the same combinatorics have the same renormalization period and the “same position” of the little Julia sets $f^k(K(g))$ in $\hat{\mathbb{C}}$ (the rigorous definition is based on the notion of Thurston’s equivalence from Holomorphic Dynamics).

Theorem 1 (Douady and Hubbard 1986). *The set of parameters c for which a quadratic map $P_c:z\mapsto z^2+c$ is renormalizable with a given combinatorics τ assemble a homeomorphic copy M^τ of the Mandelbrot set M .*

This theorem explains the presence of many little Mandelbrot sets that are observable on the computer pictures of M (see Figures 3 and 4). Moreover, the copies corresponding to the primitive renormalization originate at primitive hyperbolic components (see Holomorphic Dynamics), while the copies obtained by a satellite renormalization originate at satellite hyperbolic components attached to some

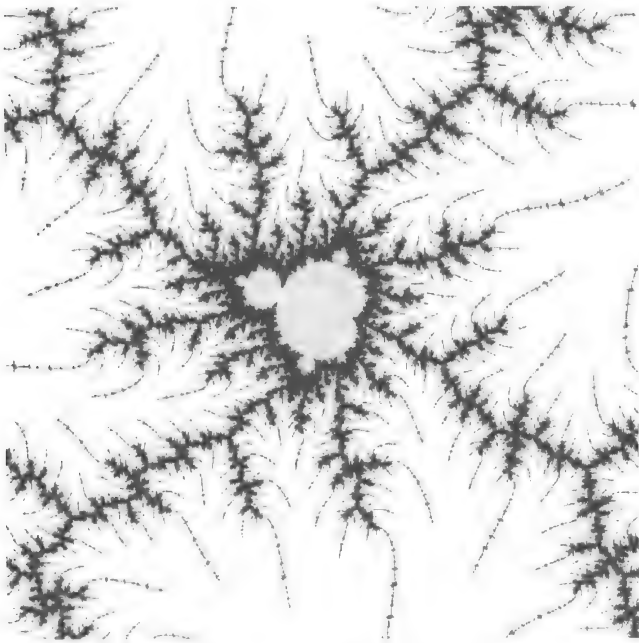


Figure 3 A primitive copy of the Mandelbrot set.

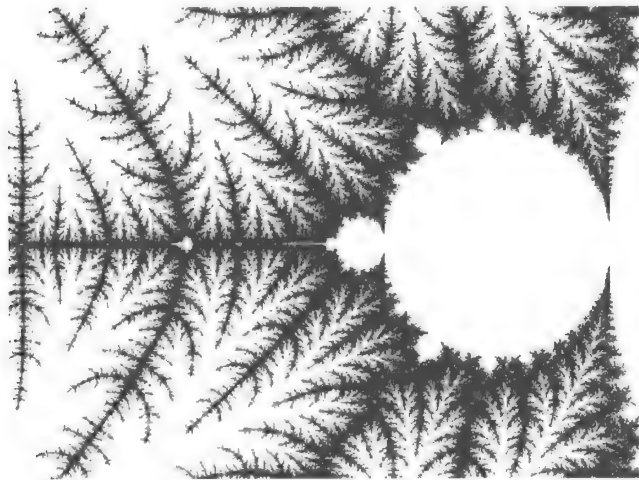


Figure 4 The satellite copy of the Mandelbrot set attached to the main cardioid at the point of doubling bifurcation.

“mother” hyperbolic component. (Satellite copies attached to the main cardioid are particularly prominent on the pictures of M .)

Given a combinatorial type τ , the set \mathcal{Q}^τ of quadratic-like germs $f \in \mathcal{Q}$ that are renormalizable with combinatorics τ (the complex renormalization strip) is the union of hybrid classes passing through the little copy M^τ . As in the real case, let us consider the set \mathcal{T}_C of all minimal combinatorial types. Then the corresponding renormalization strips \mathcal{Q}^τ are pairwise disjoint, and we can define the *full* complex renormalization operator $R: \bigcup_{\tau \in \mathcal{T}_C} \mathcal{Q}^\tau \rightarrow \mathcal{Q}$.

Renormalization Theorem

The first proof of the Renormalization Conjecture in the period-doubling case was based on rigorous

computer estimates (Lanford 1982). It followed, in the 1980s, by works of Epstein, Eckmann, Khanin, Sinai, among others, which gave a better conceptual understanding and provided proofs of many ingredients of the picture (without computer assistance).

The turning point in this development occurred when methods of holomorphic dynamics and conformal geometry were introduced into the subject (Douady and Hubbard 1985, Sullivan 1986). This led to the proof of the renormalization conjecture in the space of quadratic-like germs:

Theorem 2 (Sullivan–McMullen–Lyubich, the 1990s). *For any real combinatorics $\tau \in \mathcal{T}$, the operator R_τ has a unique fixed point f_τ in the space \mathcal{Q} . Moreover, f_τ is hyperbolic, its stable manifold $W^s(f_\tau)$ coincides with the hybrid class \mathcal{H}_c , $c = \chi(f_\tau)$, while the real slice of the unstable manifold represents an almost full family of unimodal maps.*

This result was further extended to the smooth category by de Faria, de Melo, and Pinto.

MLC, Density of Hyperbolicity, and Geometry of Feigenbaum Julia Sets

The “Mandelbrot set is locally connected” (MLC) conjecture (see Holomorphic Dynamics) is intimately related to the renormalization phenomenon. This connection was first revealed by the following result:

Theorem 3 (Yoccoz 1990, unpublished). *Let us consider a nonrenormalizable quadratic polynomial $P_c: z \mapsto z^2 + c$ with connected Julia set and both fixed points repelling. Then the Julia set $J(P_c)$ is locally connected and the Mandelbrot set is locally connected at c .*

This result was recently extended to higher-degree unicritical polynomials $z \mapsto z^d + c$ (Kahn–Lyubich, preprint 2005).

The MLC Conjecture is still open for general infinitely renormalizable parameters. However, the similar problem for the real quadratic family has been resolved. It implies the real version of the Fatou conjecture in the quadratic case (see Holomorphic Dynamics):

Theorem 4 (Lyubich 1997). *Hyperbolic maps are dense in the real quadratic family.*

This result was recently extended to higher-degree polynomials by Kozlovskii, Shen, and van Strien (preprint 2003).

Infinitely renormalizable quadratic maps of bounded combinatorial type (i.e., with bounded relative periods p_{n+1}/p_n) supply us with a rich class of fractals with very interesting geometry. These

Julia sets are “hairy” at the origin, that is, their blow-ups fill in densely the whole plane (this phenomenon is related to the universal geometry of the Feigenbaum attractors; McMullen (1996)). However, some of them have zero Lebesgue measure (Yarrington, thesis 1995) and Hausdorff dimension smaller than 2 (Avila–Lyubich, preprint 2004). It is unknown whether this happens for all of them or not (in particular, the answer is unknown for the Feigenbaum map born in the cascade of doubling bifurcations).

Regular or Stochastic Dichotomy

Stochastic Maps

An S -unimodal map f is called *stochastic* if it has an absolutely continuous invariant measure μ . In this case, f is topologically chaotic (see the section “Topological dynamics”) and μ is supported on the transitive cycle of intervals $\cup J_k$. Moreover, μ has a positive characteristic exponent,

$$\chi = \int \log |Df| d\mu > 0$$

and Lebesgue almost all orbits are equidistributed with respect to μ , that is, for Lebesgue a.e. $x \in I$,

$$\frac{1}{n} \sum \phi(f^n x) \rightarrow \int \phi d\mu$$

for any continuous function ϕ . The map $f^p|J$ is mixing with respect to μ , and in fact, is *weakly Bernoulli*.

Here are two important criteria for stochasticity:

- *Collet–Eckmann condition* (see Holomorphic Dynamics). These maps have extra strong stochastic properties, notably, the exponential decay of correlations.
- *Martens–Nowicki condition*. To state it, we need to define the principal nest of intervals, $I^0 \supset I^1 \supset \dots \ni 0$. Here $I^0 = [-\alpha, \alpha]$, where α is the fixed point with negative multiplier, and I^{n+1} is inductively defined as the component of $f^{-l_n}(I^n)$ containing 0, where l_n is the moment of first return of the orbit of 0 to I^n . Let us consider the scaling factors $\sigma_n = |I^n|/|I^{n-1}|$. If $\sum \sqrt{\sigma_n} < \infty$ then f is stochastic.

Let $\mathcal{N} \subset [-2, 1/4]$ be the set of parameters c for which the quadratic map P_c is topologically chaotic. Not every such map is stochastic. However, *the set of stochastic parameters has positive Lebesgue measure* (Jakobson 1981), and in fact,

Theorem 5 (Lyubich 2000). *For a.e. $c \in \mathcal{N}$, the map P_c satisfies the Martens–Nowicki condition, and thus, is stochastic.*

Avila and Moreira (2005) went on to prove that *for a.e. $c \in \mathcal{N}$, the map P_c is Collet–Eckmann*.

Renormalization Horseshoe

Let us consider the complexification of the renormalization operator [2],

$$R : \bigcup_{\tau \in T} \mathcal{Q}^\tau \rightarrow \mathcal{Q} \quad [3]$$

acting in the space of quadratic-like maps.

Theorem 6 (Lyubich 2002). *The “Strong Renormalization Conjecture” is valid for the operator [3].*

Let $\mathcal{I} \subset [-2, 1/4]$ be the set of parameters for which the quadratic map P_c is infinitely renormalizable. The above theorem implies that this set has *zero Lebesgue measure*. (Avila and Moreira went on to prove that $\text{HD}(\mathcal{I}) < 1$.)

Regular or Stochastic Dichotomy

Putting together Theorems 5 and 6, we obtain:

Theorem 7 *For a.e. $c \in [-2, 1/4]$, the quadratic map P_c is either regular or stochastic.*

This result gives a complete probabilistic picture of dynamics in the real quadratic family. It has been later transferred to any nondegenerate real analytic family of S -unimodal maps (Avila–Lyubich–de Melo), and further to a generic smooth family of S -unimodal maps (Avila–Moreira).

Palis has formulated a strong general conjecture (in all dimensions) asserting that a typical (from the probabilistic point of view) smooth dynamical system f has finitely many attractors supporting SRB measures (see Lyapunov Exponents and Strange Attractors) that govern the behavior of Lebesgue a.e. trajectories of f . The above results confirm the *Palis Conjecture* in the setting of S -unimodal maps.

Other Universality Classes

From a more general point of view, renormalization is an appropriately rescaled return map to a relevant piece of the phase space, viewed as an operator in some class of dynamical systems. From this point of view, most dynamical systems are “renormalizable,” and the renormalization approach often provides a deep insight into the nature of the systems in question.

Here is a partial list of classes of nonlinear systems that exhibit universality with an underlying renormalization mechanism (we provide a few

relevant names, but there are many more people who contributed to the corresponding theories):

- Holomorphic germs near indifferent equilibria (Yoccoz, Shishikura, McMullen);
- critical circle maps (Kadanoff, Feigenbaum, Rand, Lanford, Swiatek, de Faria, Yampolsky);
- non-renormalizable quadratic-like maps of Fibonacci type (Lyubich–Milnor);
- conservative two-dimensional diffeomorphisms near the point of breaking of KAM tori (MacKay, Koch); and
- dissipative Hénon-like maps (Collet–Eckmann–Koch, de Carvalho–Lyubich–Martens).

See also: Fractal Dimensions in Dynamics; Holomorphic Dynamics; Lyapunov Exponents and Strange Attractors; Multiscale Approaches.

Further Reading

Collet P and Eckmann J-P (1980) *Iterated Maps of the Interval as Dynamical Systems*. Boston: Birkhäuser.

- Cvitanović P (1984) *Universality in Chaos*. Bristol: Adam Hilger.
- Douady A and Hubbard JH (1985) On the dynamics of polynomial-like maps. *Annales Scientifiques de l'École Normale Supérieure* 18: 287–343.
- Lyubich M (1999) Feigenbaum–Coullet–Tresser universality and Milnor's hairiness conjecture. *Annals of Mathematics* 149: 319–420.
- Lyubich M (2000) Quadratic family as a qualitatively solvable model of dynamics. *Notices of the American Mathematical Society* 47(9): 1042–1052.
- McMullen C (1996) *Renormalization and 3-Manifolds Which Fiber Over the Circle*, *Annals of Math. Studies*, vol. 135. Princeton: Princeton University Press.
- de Melo W and van Strien S (1993) *One-Dimensional Dynamics*. Berlin: Springer.
- Sullivan D (1993) Linking the universalities of Milnor–Thurston, Feigenbaum and Ahlfors–Bers. In: *Topological Methods in Modern Mathematics*, The Proceedings of Symposium held in honor of John Milnor's 60th Birthday, SUNY at Stony Brook, 1991, pp. 543–564. Houston, TX: Publish or Perish.
- Vul EB, Sinai YaG, and Khanin KM (1984) Feigenbaum universality and the thermodynamical formalism. *Russian Mathematical Surveys* 39: 1–40.

V

Variational Methods in Turbulence

F H Busse, Universität Bayreuth, Bayreuth, Germany

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The problem of fluid turbulence is commonly regarded as one of the most challenging problems of theoretical physics and mathematics. There is general agreement that the Navier–Stokes equations (NSEs) provide a satisfactory basis for the description of turbulent motions of homogeneous Newtonian fluids such as gases and most liquids. But the difficulty of generating solutions of these equations for high-Reynolds-number flows has prevented accurate answers to simple questions such as the question of the discharge of turbulent pipe flow as a function of the pressure head or the question of the heat transport by turbulent convection in a fluid layer heated from below. In view of this difficulty, it has become an attractive idea to obtain rigorous bounds on turbulent transports. Variational methods have played an important role in the derivation of such bounds.

There is another motivation for the use of variational methods for the understanding of turbulent fluid systems. Experimenters have sometimes noted the tendency of turbulent flows to maximize transports under given external conditions. In his pioneering paper, Howard (1963) mentions that the Malkus hypothesis of a maximum heat transport by thermal convection had motivated him to derive upper bounds through the use of variational methods. The techniques developed by Howard have later been applied to other kinds of turbulent transports by Busse. While relatively simple ordinary differential equations are obtained when the equation of continuity is not imposed as a constraint, the Euler–Lagrange equations for a stationary value of the variational functional lead to nonlinear partial differential equations when solenoidal extremalizing vector fields are required. Nevertheless, using boundary layer methods one can derive approximate analytical solutions even in the limit of asymptotically large Rayleigh and Reynolds numbers (Busse 1969, 1978).

In the following, we shall first discuss the energy method which provides necessary conditions for the existence of turbulent solutions of the underlying equations and then turn to the problem of upper bounds for the turbulent momentum transport in the plane Couette flow configuration as a particular example. The properties and physical relevance of the extremalizing vector fields will be discussed in a final section.

Energy Method

For simplicity, we consider the NSEs for a homogeneous incompressible fluid with a constant kinematic viscosity ν in an arbitrary fixed domain \mathcal{D} . Using the diameter d of the domain as length scale and d^2/ν as timescale, we can write the NSEs of motion in dimensionless form,

$$\frac{\partial}{\partial t} \mathbf{v} + \mathbf{v} \cdot \nabla \mathbf{v} = -\nabla p + \mathbf{f} + \nabla^2 \mathbf{v} \quad [1a]$$

$$\nabla \cdot \mathbf{v} = 0 \quad [1b]$$

where \mathbf{f} denotes some given steady distribution of a force density. On the boundary $\partial\mathcal{D}$ of the domain \mathcal{D} , steady velocities parallel to the boundary may be specified. We assume that the basic steady solution of the problem is given by $\mathbf{v}_s = Re \hat{\mathbf{v}}$ where the average of $(\hat{\mathbf{v}})^2/2$ over the domain \mathcal{D} (indicated by angular brackets) is unity, $\langle |\hat{\mathbf{v}}|^2 \rangle = 2$. Any velocity field \mathbf{v}_t different from \mathbf{v}_s , that is, with $\mathbf{u} \equiv \mathbf{v}_t - \mathbf{v}_s \neq 0$, must obey the equations

$$\frac{\partial}{\partial t} \mathbf{u} + \mathbf{v}_s \cdot \nabla \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{v}_s + \mathbf{u} \cdot \nabla \mathbf{u} = -\nabla \tilde{p} + \nabla^2 \mathbf{u} \quad [2a]$$

$$\nabla \cdot \mathbf{u} = 0 \quad [2b]$$

together with the homogeneous boundary conditions for \mathbf{u} on $\partial\mathcal{D}$. By multiplying eqn [2a] by \mathbf{u} and averaging the result over the domain \mathcal{D} we obtain the relationship

$$\frac{1}{2} \frac{d}{dt} \langle \mathbf{u} \cdot \mathbf{u} \rangle = -\langle |\nabla \mathbf{u}|^2 \rangle - Re \langle \mathbf{u} \cdot (\mathbf{u} \cdot \nabla) \hat{\mathbf{v}} \rangle \quad [3]$$

where the vanishing of \mathbf{u} on $\partial\mathcal{D}$ and equations such as

$$\begin{aligned}\langle \mathbf{u} \cdot (\mathbf{v}_s \cdot \nabla) \mathbf{u} \rangle &= \frac{1}{2} \langle \mathbf{v}_s \cdot \nabla \mathbf{u} \cdot \mathbf{u} \rangle \\ &= \frac{1}{2} \langle \nabla \cdot (\mathbf{v}_s \mathbf{u} \cdot \mathbf{u}) \rangle = 0\end{aligned}$$

have been used to prove that the terms $\mathbf{v}_s \cdot \nabla \mathbf{u}$, $\mathbf{u} \cdot \nabla \mathbf{u}$ and $\nabla \tilde{p}$ do not enter the balance [3]. This balance is called the Reynolds–Orr energy equation and is the basis for the application of the energy method. The lowest value Re for which the right-hand side of [3] is non-negative is called the energy Reynolds number Re_E . For $Re < Re_E$ the steady solution \mathbf{v}_s is absolutely stable and the energy of any disturbance \mathbf{u} must decay exponentially in time. $Re > Re_E$ is a necessary condition for the existence of a persistent turbulent state of fluid flow. Re_E is determined as the solution of the variational problem:

For a given flow $\hat{\mathbf{v}}$ in \mathcal{D} find the minimum Re_E of the functional

$$\mathcal{R}_E \equiv \frac{\langle |\nabla \tilde{\mathbf{u}}|^2 \rangle}{\langle -\tilde{\mathbf{u}} \cdot (\tilde{\mathbf{u}} \cdot \nabla) \hat{\mathbf{v}} \rangle} \tag{4}$$

among all vector fields $\tilde{\mathbf{u}}$ which satisfy the conditions $\nabla \cdot \tilde{\mathbf{u}} = 0$ in \mathcal{D} , $\tilde{\mathbf{u}} = 0$ on $\partial\mathcal{D}$, and $\langle \tilde{\mathbf{u}} \cdot (\tilde{\mathbf{u}} \cdot \nabla) \hat{\mathbf{v}} \rangle < 0$.

For $Re \geq Re_E$ there will exist at least one vector field \mathbf{u} , namely the minimizing solution $\tilde{\mathbf{u}}$ of the variational problem [4], the energy of which does not decay, at least not initially. In the derivation of the Euler–Lagrange equations as necessary conditions for stationary values of the variational functional [4],

$$\frac{1}{2} G(\tilde{\mathbf{u}}_\kappa \partial_\kappa \hat{\mathbf{v}}_i + \tilde{\mathbf{u}}_\kappa \partial_i \hat{\mathbf{v}}_\kappa) = -\partial_i \tilde{\pi} + \partial_\kappa \partial_\kappa \tilde{\mathbf{u}}_i \tag{5a}$$

$$\partial_\kappa \tilde{\mathbf{u}}_\kappa = 0 \tag{5b}$$

the constraint $\nabla \cdot \tilde{\mathbf{u}} = 0$ has been taken into account through the Lagrange multiplying function $\tilde{\pi}$. G is a stationary value of the functional [4] and in general there exist many of those which are determined as eigenvalues of the linear boundary value problem [5] together with its boundary condition $\tilde{\mathbf{u}}_i = 0$ on $\partial\mathcal{D}$. Only the infimum of all G provides the energy Reynolds number Re_E . Many details on the energy method can be found in Joseph’s book (1976). Here we just wish to remark that the Reynolds–Orr balance [3] remains valid when the problem is considered in a system rotating with a constant angular velocity Ω_D since the Coriolis force does not contribute to the energy balance [3]. The values of Re_E are usually much smaller than the critical values Re_c for the onset of infinitesimal disturbances as can be seen from Table 1. Here the experimentally determined values

Table 1 Reynolds numbers for shear flows

	Re_E	Re_G (from exp.)	Re_c
Plane Couette flow	82.6	≈ 1300	∞
Poiseuille flow (channel flow)	99.2 ^a	$\approx 2000^a$	5772 ^a
Hagen–Poiseuille flow (pipe flow)	81.5 ^a	$\approx 2100^a$	∞
Circular Couette flow with $\Omega_D = Re_E/2$	82.6	≈ 82.6	82.6

^aThe maximum velocity and the channel width d (radius d in the case of pipe flow) have been used in definition of Re .

Re_G for the instability of the basic flow state have also been listed. A unique situation occurs in the small gap limit of the Taylor–Couette system where Re_E and Re_c coincide for a special value of the dimensionless mean rotation rate Ω_D (Busse 2002).

Variational Problem for Turbulent Momentum Transport

In order to introduce the variational method for bounds on turbulent transports we consider the simplest configuration for which a nontrivial solution of the NSEs of motion exists: the configuration of plane Couette flow (Figure 1). The Reynolds number is defined in this case in terms of the constant relative motion $U_0 \mathbf{i}$ between the plates, $Re = U_0 d / \nu$, where \mathbf{i} is the unit vector parallel to the plates and ν is the kinematic viscosity of the fluid. Using the distance d between the plates as length scale and d^2 / ν as timescale, the basic equations can be written in the form

$$\frac{\partial}{\partial t} \mathbf{v} + \mathbf{v} \cdot \nabla \mathbf{v} = -\nabla p + \nabla^2 \mathbf{v} \tag{6}$$

$$\nabla \cdot \mathbf{v} = 0 \tag{7}$$

We use a Cartesian system of coordinates with the x , z -coordinates in the directions of \mathbf{i} and \mathbf{k} ,

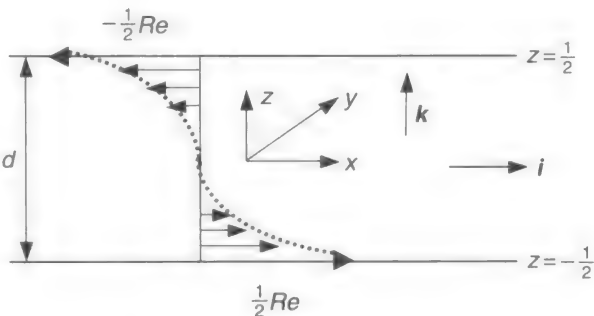


Figure 1 Geometrical configuration of the plane Couette flow problem.

respectively, where \mathbf{k} is the unit vector normal to the plates such that the boundary conditions are given by

$$\mathbf{v} = \mp \frac{1}{2} Re \mathbf{i} \quad \text{at } z = \pm \frac{1}{2} \quad [8]$$

After separating the velocity field \mathbf{v} into its mean and fluctuating parts, $\mathbf{v} = U + \tilde{\mathbf{v}}$ with $\bar{\mathbf{v}} = U$, $\tilde{\mathbf{v}} = 0$, where the bar denotes the average over planes $z = \text{const.}$, we obtain by multiplying eqn [6] by $\tilde{\mathbf{v}}$ and averaging it over the entire fluid layer (indicated by angular brackets)

$$\frac{1}{2} \frac{d}{dt} \langle |\tilde{\mathbf{v}}|^2 \rangle = - \left\langle \overline{\mathbf{u}\mathbf{w}} \cdot \frac{\partial}{\partial z} U \right\rangle - \langle |\nabla \tilde{\mathbf{v}}|^2 \rangle \quad [9]$$

Here \mathbf{u} denotes the component of $\tilde{\mathbf{v}}$ perpendicular to \mathbf{k} and w is its z -component. We define fluid turbulence under stationary conditions by the property that quantities averaged over planes $z = \text{const.}$ are time independent. Accordingly, the equation for the mean flow U can be integrated to yield

$$\frac{d}{dz} U = \overline{w\mathbf{u}} - \langle w\mathbf{u} \rangle - Re \mathbf{i} \quad [10]$$

where the boundary condition [8] has been employed. With this relationship, U can be eliminated from the problem and the energy balance

$$\langle |\nabla \mathbf{u}|^2 \rangle + \langle |\overline{w\mathbf{u}} - \langle w\mathbf{u} \rangle|^2 \rangle = Re \langle u_x w \rangle \quad [11]$$

is obtained where the identity $\langle \overline{w\mathbf{u}}^2 \rangle - \langle w\mathbf{u} \rangle^2 = \langle |\overline{w\mathbf{u}} - \langle w\mathbf{u} \rangle|^2 \rangle$ has been used.

Since the momentum transport in the x -direction between the moving rigid plates is described by $M = -dU_x/dz|_{z=\pm 1/2} = \langle u_x w \rangle + Re$, we can conclude immediately that the momentum transport by turbulent flow always exceeds the corresponding laminar value because $\langle u_x w \rangle$ is positive according to the relationship [11]. Since a lower bound on M thus exists, an upper bound μ on $\langle u_x w \rangle$ as a function of Re is of primary interest. Following Howard (1963), it can be shown that $\mu(Re)$ is a monotonous function and it is therefor equivalent to ask for a lower bound R of Re at a given value μ of $\langle u_x w \rangle$. We are thus led to the following formulation of the variational problem:

Find the minimum $R(\mu)$ of the functional

$$\mathcal{R}(\tilde{\mathbf{v}}, \mu) \equiv \frac{\langle |\nabla \tilde{\mathbf{v}}|^2 \rangle}{\langle \tilde{u}_x \tilde{w} \rangle} + \mu \frac{\langle |\tilde{u}\tilde{w} - \langle \tilde{u}\tilde{w} \rangle|^2 \rangle}{\langle \tilde{u}_x \tilde{w} \rangle^2} \quad [12]$$

among all solenoidal vector fields $\tilde{\mathbf{v}} \equiv \tilde{\mathbf{u}} + k\tilde{w}$ (with $\tilde{\mathbf{u}} \cdot \mathbf{k} = 0$) that satisfy the boundary condition $\tilde{\mathbf{v}} = 0$ at $z = \pm 1/2$ and the condition $\langle \tilde{u}_x \tilde{w} \rangle > 0$.

The Euler-Lagrange equations as necessary conditions for an extremal value of the functional are given by

$$\tilde{w} \frac{d}{dz} U^* + k\tilde{\mathbf{u}} \cdot \frac{d}{dz} U^* = -\nabla \pi + \nabla^2 \tilde{\mathbf{v}} \quad [13]$$

$$\nabla \cdot \tilde{\mathbf{v}} = 0 \quad [14]$$

where dU^*/dz is defined by

$$\frac{d}{dz} U^* = \overline{\tilde{u}\tilde{w}} - \langle \tilde{u}\tilde{w} \rangle - i \left(R - \frac{\langle |\nabla \tilde{\mathbf{v}}|^2 \rangle}{2\langle \tilde{u}_x \tilde{w} \rangle} \right) \quad [15]$$

and where $\mu = \langle \tilde{u}_x \tilde{w} \rangle$ has been set. When eqns [13]–[15] are compared with the equations for $\tilde{\mathbf{v}}$ and for U , a strong similarity can be noticed. The variational problem does not exhibit any time dependence, but the Euler-Lagrange equations may still be regarded as the symmetric analogue of the NSEs for steady flow.

Upper Bounds on the Turbulent Momentum Transport

A simple analytical solution of the variational problem can be obtained when the constraint $\nabla \cdot \tilde{\mathbf{v}} = 0$ is dropped. In that case it is evident that the minimum of the functional [12] is reached when $\tilde{\mathbf{v}}$ is independent of x , y , and when $\tilde{u}_x = \tilde{w} = f(z)$ holds. The Euler-Lagrange equations then assume the form of an ordinary differential equation,

$$f'' = [\mu(f^2/\langle f^2 \rangle - 1) - R + \langle f'^2 \rangle/\langle f^2 \rangle]f \quad [16]$$

Since the variational functional [12] is homogeneous in $\tilde{\mathbf{v}}$, we are free to use a normalization condition for which we choose $\max[f(z)] = 1$. Multiplication of eqn [16] by f' and integration yield

$$f'^2 = \frac{\mu}{2k^2\langle f^2 \rangle} (1 - k^2 f^2)(1 - f^2) \quad [17]$$

with $k^2 = \mu/[2(R + \mu)\langle f^2 \rangle - 2\langle f'^2 \rangle - \mu]$

This equation can be solved in terms of elliptical integrals. The minimum $R(\mu)$ is determined by the relationships

$$R = \frac{8}{3} [K^2(1 + k^2) + K^3/D - 3k^2KD] \quad [18]$$

$$\mu = 8k^2KD$$

where $D(k)$ and $K(k)$ are the complete elliptical integrals usually labeled by these letters. For details, see the analysis by Howard (1963) of an analogous problem. In the asymptotic case of large

Reynolds numbers, relationships [18] yield the upper bound

$$\mu(Re) = \frac{8}{128} Re^2 \quad [19]$$

In solving the full eqns [13]–[15], it is convenient to eliminate eqn [14] through the general representation of the solenoidal vector field \tilde{v} ,

$$\tilde{v} = \nabla \times (\nabla \phi \times \mathbf{k}) + \nabla \psi \times \mathbf{k} \quad [20]$$

We assume that the minimizing vector field \tilde{v} does not depend on x , although a rigorous proof for this property can be given only for small values of μ . Introducing the notations $\theta \equiv \partial \psi / \partial y$ and $w \equiv -\partial^2 \phi / \partial y^2$ we are thus led to the general ansatz

$$w = w^{(N)} \equiv \sum_{n=1}^N \alpha_n^{-2} w_n(z) \phi_n(y) \quad [21a]$$

$$\theta = \theta^{(N)} \equiv \sum_{n=1}^N \theta_n(z) \phi_n(y) \quad [21b]$$

where N may tend to infinity and the functions $\phi_n(y)$ satisfy the equation

$$\frac{\partial^2}{\partial y^2} \phi_n = -\alpha_n^2 \phi_n \quad [22]$$

In the following, it will be assumed that the positive wavenumbers α_n are ordered according to their size, $\alpha_{n-1} < \alpha_n < \alpha_{n+1}$. The solutions of the form [21] of the Euler–Lagrange equations exhibit a boundary layer structure for large μ as sketched in Figure 2. Accordingly, the N - α solutions are characterized by a hierarchy of N boundary layers at each plate and provide the upper bound sequentially with increasing μ starting with $N = 1$. The extremalizing vector fields thus exhibit a bifurcation structure similar to that found in many cases of the transition to turbulence. The thicknesses of the boundary layers decrease with increasing μ and their ratio from one layer to the next approaches the factor 4 as indicated in Figure 3. The typical scale of motion increases linearly with

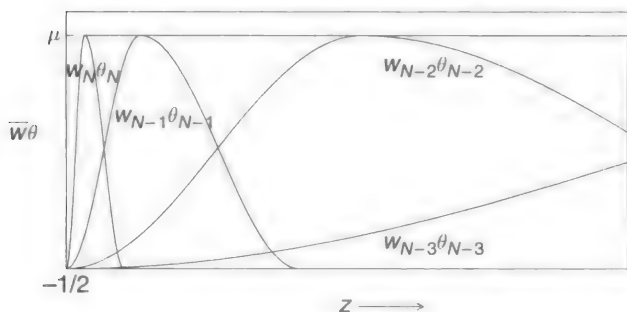


Figure 2 Qualitative sketch of the boundary layer structure of the extremalizing N - α solution.

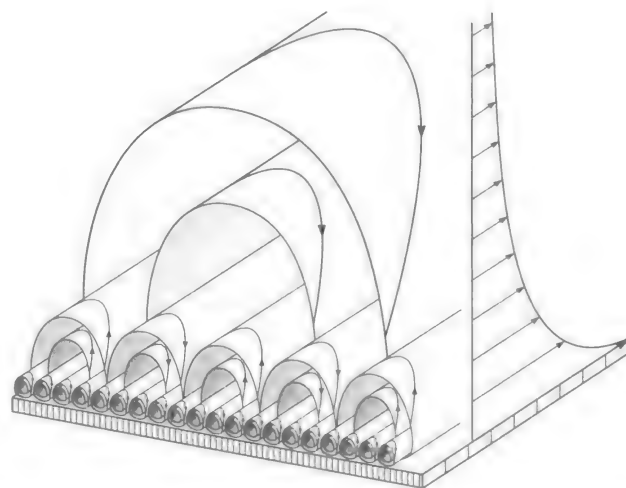


Figure 3 Qualitative sketch of the nested boundary layers that characterize the vector field of maximum transport. The profile of the mean shear is shown on the right side.

distance from the wall as assumed in Prandtl's mixing-length theory. But the discreteness of the scales reflects the fact that effective transports require preferred scales. Asymptotically, the upper bound for the momentum transport approaches

$$\mu(Re) = 0.010 Re^2 \quad [23]$$

which represents a significant improvement over the relationship [19]. Nevertheless, the upper bound still exceeds the measured values of the momentum transport by more than a factor 10.

Discussion

Bounds like those for the momentum transport have been obtained for many other kinds of turbulent transports. For details we refer to the review articles listed below. Usually, the formulation of the upper bound problem requires that the external conditions are homogeneous in two spatial dimensions such that a separation of the turbulent velocity, temperature, or magnetic fields into mean and fluctuating parts is possible. In this respect, the variational methods for upper bounds are more restricted than those used for determination of the energy Reynolds number Re_E . The latter problem, incidentally, corresponds to the limit $\mu \rightarrow 0$ of variational problems of the type [12] as can be seen from a comparison with expression [4].

In recent years, the background field method has been introduced by Doering and Constantin (1994) as an alternative way for obtaining bounds on properties of turbulent flows. When optimized, it becomes equivalent to the variational method discussed in this article as has been demonstrated by Kerswell (1998). The fact that not optimized bounds can be obtained

relatively easily emphasizes the point that the extremalizing vector fields are the most interesting aspect of the variational problems. They often exhibit similarities with the observed turbulent velocity fields, in particular as far as the mean flows are concerned. In the case of convection in a layer heated from below, the transition of the bound from the $1-\alpha$ solution to the $2-\alpha$ solution corresponds closely to the experimentally observed transition from convection rolls to bimodal convection (Busse 1969).

The close similarities between variational functionals for rather different physical systems suggest corresponding similarities between the respective turbulent fields. For example, the analogy between the fluctuating component of the temperature in turbulent convection and the streamwise component of the fluctuating velocity field in shear flow turbulence has been demonstrated and employed in a theory of the atmospheric boundary layer (Busse 1978). Better bounds and more physically realistic properties of the extremalizing vector fields can be expected when additional constraints are imposed. For example, the energy balances for poloidal and toroidal components of the velocity field can be applied separately. But these developments are still in their initial stages.

See also: Bifurcations in Fluid Dynamics; Fluid Mechanics: Numerical Methods; Turbulence Theories.

Further Reading

- Busse FH (1969) On Howard's upper bound for heat transport by turbulent convection. *Journal of Fluid Mechanics* 37: 457–477.
- Busse FH (1978) The optimum theory of turbulence. *Advances in Applied Mechanics* 18: 77–121.
- Busse FH (2002) The problem of turbulence and the manifold of asymptotic solutions of the Navier–Stokes equations. In: Oberlack M and Busse FH (eds.) *Theories of Turbulence*, pp. 77–121. Wien: Springer.
- Doering CR and Constantin P (1994) Variational bounds on energy dissipation in incompressible flows: shear flow. *Physical Review E* 49: 4087–4099.
- Howard LN (1963) Heat transport by turbulent convection. *Journal of Fluid Mechanics* 17: 405–432.
- Howard LN (1972) Bounds on flow quantities. *Annual Review of Fluid Mechanics* 4: 473–494.
- Joseph DD (1976) *Stability of fluid motions*. vol. 1. Berlin: Springer.
- Kerswell RR (1998) Unification of variational principles for turbulent shear flows: the background method of Doering–Constantin and Howard–Busse's mean-fluctuation formulation. *Physica D* 121: 175–192.

Variational Techniques for Ginzburg–Landau Energies

S Serfaty, New York University, New York, NY, USA

© 2006 Elsevier Ltd. All rights reserved.

Ginzburg–Landau-type problems are variational problems which consider a Dirichlet-type energy posed on complex-valued functions, penalized by a potential term which has a well in the unit circle of the complex plane. The denomination comes from the physical model of superconductivity of Ginzburg and Landau. They are phase-transition-type models in the sense that they describe the state of the material according to different “phases” which can coexist in a sample and be separated by various types of interfaces. We start by presenting the physical model (readers familiar with it may wish to skip the next two sections and go straight to the section “The simplified model”).

Introduction to the Ginzburg–Landau Model

The Ginzburg–Landau model was introduced by Ginzburg and Landau in the 1950s as a phenomenological model to describe superconductivity, and was later justified as a limit of the quantum

BCS theory of Bardeen–Cooper–Schrieffer. It is a model of great importance and recognition in physics (with several Nobel prizes awarded for it: Landau, Ginzburg, Abrikosov). In addition to its importance in the modeling of superconductivity, the Ginzburg–Landau model turns out to be mathematically extremely close to the Gross–Pitaevskii model for superfluidity, and models for rotating Bose–Einstein condensates, which all have in common the appearance of topological defects called “vortices.”

Superconductivity, which was discovered in 1911 by Kammerling Ohnes, consists in the complete loss of resistivity of certain metals and alloys at very low temperatures: the two most striking consequences of it being the possibility of permanent superconducting currents and the particular behavior that an external magnetic field applied to the sample gets expelled from the material and can generate vortices, through which it penetrates the sample.

The Energy Functional

After a series of dimension reductions, the Ginzburg–Landau model describes the state of the superconducting sample occupying a region Ω and submitted to the external magnetic field h_{ex} ,

below the critical temperature, through its Gibbs energy:

$$G_\varepsilon(\psi, A) = \frac{1}{2} \int_\Omega |\nabla_A \psi|^2 + \frac{(1 - |\psi|^2)^2}{2\varepsilon^2} + \frac{1}{2} \int_{\mathbb{R}^3} |\operatorname{curl} A - h_{\text{ex}}|^2 \quad [1]$$

In this expression, the first unknown ψ is the “order parameter” in physics. It is a complex-valued condensed wave function, indicating the local state of the material, or the phase (in the Landau theory approach of phase transitions): $|\psi|^2$ is the density of the “Cooper pairs” of superconducting electrons explaining superconductivity in the BCS approach. With our normalization $|\psi| \leq 1$ and where $|\psi| \sim 1$ the material is in the superconducting phase, while where $|\psi| \sim 0$, it is in the normal phase (i.e., behaves like a normal conductor), the two phases being able to coexist in the sample.

The second unknown A is the electromagnetic vector potential of the magnetic field, a function from Ω to \mathbb{R}^3 . The induced magnetic field in the sample is deduced by $h = \operatorname{curl} A$. The notation ∇_A denotes the covariant derivative $\nabla - iA$. The superconducting current is the vector j of components

$$j_k = \langle i\psi, (\nabla_A)_k \psi \rangle \quad [2]$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product in \mathbb{C} identified with \mathbb{R}^2 .

Finally, the parameter ε is the inverse of the “Ginzburg–Landau parameter” κ , a dimensionless parameter (ratio of the penetration depth and the coherence length) depending on the material only.

Most variational studies of Ginzburg–Landau focus on the regime of large κ or small ε , corresponding to “extreme type-II” superconductors, also called the London limit. In this limit, the potential term acts as a singular perturbation, and the characteristic size of the vortices is $\varepsilon \rightarrow 0$; vortices become line-like topological singularities, which makes it easier to extract and describe them.

This model is a $U(1)$ -gauge theory, that is, it is invariant under the gauge transformations:

$$\begin{aligned} \psi &\mapsto \psi e^{i\Phi} \\ A &\mapsto A + \nabla \Phi \end{aligned} \quad [3]$$

where Φ is a smooth real-valued function. The physically relevant quantities are those that are gauge invariant, such as the energy G_ε , $|\psi|$, h , and the superconducting current j .

For more on the model, we refer to the physics literature (e.g., DeGennes (1966) and Tinkham (1996)).

Reductions of the Model

The goal of variational studies of the Ginzburg–Landau model is to relate the energy to the vortices and the applied field. In three dimensions (3D), vortices are filaments, or lines of zeros of the order parameter ψ , around which ψ has a nonzero winding number. These are quite delicate to describe in 3D (we will mention some results below), so a simplification that is commonly made consists in reducing to a two-dimensional model.

When reducing to 2D, one assumes that everything is independent of the vertical direction, and that the applied magnetic field is also vertical. The domain Ω is then a two-dimensional, bounded and (for simplicity) simply connected open set, which is the horizontal section of an infinite vertical cylinder. One can also imagine it represents a thin film.

In 2D, the energy is written the same way:

$$G_\varepsilon(\psi, A) = \frac{1}{2} \int_\Omega |\nabla_A \psi|^2 + \frac{(1 - |\psi|^2)^2}{2\varepsilon^2} + |\operatorname{curl} A - h_{\text{ex}}|^2 \quad [4]$$

where this time A is \mathbb{R}^2 -valued, and the induced magnetic field $h = \operatorname{curl} A = \partial_1 A_2 - \partial_2 A_1$ is now a real-valued function, which can be taken to be equal to h_{ex} (now a real positive number) in $\mathbb{R}^2 \setminus \Omega$.

The stationary states of the system are the critical points of G_ε , or the solutions of the Ginzburg–Landau equations:

$$\begin{aligned} -(\nabla_A)^2 \psi &= \frac{1}{\varepsilon^2} \psi(1 - |\psi|^2) && \text{in } \Omega \\ -\nabla^\perp h &= \langle i\psi, \nabla_A \psi \rangle && \text{in } \Omega \\ h &= h_{\text{ex}} && \text{on } \partial\Omega \\ \nabla_A \psi \cdot \nu &= 0 && \text{on } \partial\Omega \end{aligned} \quad [5]$$

where ∇^\perp denotes $(-\partial_{x_2}, \partial_{x_1})$.

A common simplification consists in suppressing the magnetic field, and thus in studying the simplified energy

$$E_\varepsilon(u) = \frac{1}{2} \int_\Omega |\nabla u|^2 + \frac{(1 - |u|^2)^2}{2\varepsilon^2} \quad [6]$$

where the order parameter is commonly denoted by u , and is still complex valued. This energy, which can be seen as a complex analog of the real-valued Allen–Cahn model of phase transitions, has been extensively studied, especially since the work of Bethuel–Brezis–Hélein, where the domain Ω is assumed to be two dimensional and simply connected. The higher-dimensional case has also been considered.

Vortices and Critical Fields

We now need to explain more precisely what a vortex is. In two dimensions, a vortex is an object centered at an isolated zero of u (or ψ), around which the phase of u has a nonzero winding number called the “degree of the vortex.” It is the simplest example of a topological defect. If the zero is located at x_0 , the winding number or degree is the integer that can be computed by

$$\frac{1}{2\pi} \int_{\partial B(x_0, r)} \frac{\partial \varphi}{\partial \tau} = d \in \mathbb{Z} \quad [7]$$

where r is small enough, and φ is the phase of u , that is, u can be written $u = |u|e^{i\varphi}$. For example, the phase $\varphi = d\theta$, where θ is the polar angle centered at x_0 , yields a vortex of degree d . Observe that the phase φ is not a well-defined function, it is multivalued (and defined up to 2π); however, we have the important relation

$$\text{curl } \nabla \varphi = 2\pi \sum_i d_i \delta_{a_i} \quad [8]$$

where the a_i ’s are the zeros of u , d_i ’s the associated degrees, and δ_x denotes the Dirac mass at x .

When ε is small, it is clear from [4] or [6] that $|u|$ prefers to be close to 1, and a scaling argument hints that $|u|$ is different from 1 in regions of characteristic size ε . Of course this is an intuitive picture and several mathematical notions are used to describe the vortices.

Vortices appear due to the applied field h_{ex} . For type-II superconductors there are essentially three critical fields, H_{c_1} , H_{c_2} , H_{c_3} , critical values of h_{ex} for which phase transitions occur. For $h_{\text{ex}} \leq H_{c_1} = O(|\log \varepsilon|)$, there are no vortices and the superconductor is in the superconducting phase $|\psi| \simeq 1$ everywhere. At H_{c_1} the first vortices appear, and their number increases as h_{ex} is raised. When they become numerous they tend to arrange in triangular lattices called Abrikosov lattices, as observed in experiments and predicted by Abrikosov from the Ginzburg–Landau model, in a very influential work. At the second critical field $H_{c_1} = O(1/\varepsilon^2)$ bulk superconductivity is destroyed, and surface superconductivity remains until $H_{c_3} = O(1/\varepsilon^2)$, the third critical field, above which $\psi \equiv 0$ and the material is normal.

Issues and Methods

The variational approach to Ginzburg–Landau consists in expressing the energy in terms of reduced quantities or objects, in particular in terms of the vortices. This requires to develop mathematical tools to describe and characterize the vortices (in particular give some suitable definitions of a “vortex structure”

for a given u or ψ), and estimate precisely the energetic cost of each vortex and of their interaction. This allows us to obtain results of variational convergence of the energy $G_\varepsilon, E_\varepsilon$ (or their variants), that is, to derive Γ -limits, or “reduced problems” posed in terms of the vortices, which are easier to minimize than the original ones. These limits depend on the regime of applied field, and allow to characterization of, in turn, the critical fields, and the optimal repartition and number of the vortices, if any.

Variational methods also serve to solve some inverse problems, that is, to prove the existence of solutions of the equation which have some given properties, such as a given repartition of vortices, through local minimization procedures, or the use of topological methods based on investigating the topology of the energy levels.

Nonvariational approaches of Ginzburg–Landau are also very useful, in particular to identify the profiles of the solutions, to describe vortices of nonminimizing critical points, or to perform a bifurcation analysis around the normal solution at H_{c_3} .

The Simplified Model

We first present the variational study of E_ε [6] in dimension 2, together with the mathematical tools used for both [6] and [4]. We will restrict to the asymptotics $\varepsilon \rightarrow 0$, since this is the situation where the most results are known.

Let us present informally the essential ingredients of the analysis.

Tracing the Vortices

The easiest way to trace the vortices is to use the current $\langle iu, \nabla u \rangle$ (or the “superconducting current” $j = \langle i\psi, \nabla_A \psi \rangle$ for the case with magnetic field). Here we recall $\langle \cdot, \cdot \rangle$ denotes the scalar product in \mathbb{C} as identified with \mathbb{R}^2 , that is, $\langle iu, \nabla u \rangle = \langle u \times \partial_1 u, u \times \partial_2 u \rangle$ with \times the vector product in \mathbb{R}^2 .

The curl of the current is the vorticity of the map u , exactly like in fluid mechanics. Writing $u = \rho e^{i\varphi}$ we have (at least formally) $\langle iu, \nabla u \rangle = \rho^2 \nabla \varphi$ and since $\rho = |u|$ is close to 1 (other than in the small vortex regions), we have the approximation

$$\begin{aligned} \text{curl } \langle iu, \nabla u \rangle &= \text{curl } (\rho^2 \nabla \varphi) \simeq \text{curl } \nabla \varphi \\ &= 2\pi \sum_i d_i \delta_{a_i} \end{aligned} \quad [9]$$

where the a_i ’s are the zeros of u (or its vortices) and the d_i ’s their degrees, or

$$\text{curl } \langle i\psi, \nabla_A \psi \rangle + \text{curl } A \simeq \text{curl } \nabla \varphi$$

in the case with magnetic field. This can be made rigorous (see Jerrard and Soner (2002) and Sandier and Serfaty (to appear)), that is, one can express that

$$\operatorname{curl} \langle iu, \nabla u \rangle - 2\pi \sum_i d_i \delta_{a_i} \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0 \quad [10]$$

(or respectively $\operatorname{curl} \langle i\psi, \nabla_A \psi \rangle + \operatorname{curl} A - 2\pi \sum_i d_i \delta_{a_i} \rightarrow 0$) in some weak functional norm, thus giving a rigorous use of [8]. The quantity

$$\mu(u) = \operatorname{curl} \langle iu, \nabla u \rangle \quad [11]$$

or

$$\mu(\psi, A) = \operatorname{curl} \langle i\psi, \nabla_A \psi \rangle + \operatorname{curl} A = \operatorname{curl} j + h \quad [12]$$

in the case with magnetic field, will thus be called the vorticity and be used to trace the vortices, in this limit $\varepsilon \rightarrow 0$. The relation

$$\mu - 2\pi \sum_i d_i \delta_{a_i} \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0 \quad [13]$$

states that it is close to being a measure.

This is also called the Jacobian determinant if written (with differential forms) $Ju = d\langle iu, du \rangle = \langle i du, du \rangle = 2(u_{x_1} \times u_{x_2}) dx_1 \wedge dx_2$, and under this form it can be used in higher dimensions.

The Cost of Each Vortex

Here we investigate informally the cost of a vortex of degree d . We know already that the characteristic length scale of variation of u is ε , and that $(1 - |u|^2)^2$ is strongly penalized. Thus, we may expect that $|u|$ is close to 1 at a distance $\gg \varepsilon$ of the zeros. Assuming that x_0 is a zero of u , and taking formally $|u| = 1$ for $|x - x_0| \geq \varepsilon$, we may write $u = e^{i\varphi}$ and $|\nabla u| = |\nabla \varphi|$ for $|x - x_0| \geq \varepsilon$.

Then, we have

$$\begin{aligned} & \frac{1}{2} \int_{R \geq |x - x_0| \geq \varepsilon} |\nabla u|^2 \\ & \geq \frac{1}{2} \int_{\varepsilon}^R \left(\int_{\partial B(x_0, r)} \left| \frac{\partial \varphi}{\partial \tau} \right|^2 \right) dr \\ & \geq \frac{1}{2} \int_{\varepsilon}^R \left(\left(\int_{\partial B(x_0, r)} \frac{\partial \varphi}{\partial \tau} \right)^2 \frac{1}{2\pi r} \right) dr \quad [14] \end{aligned}$$

$$\geq \frac{1}{2} \frac{4\pi^2 d^2}{2\pi} \int_{\varepsilon}^R \frac{dr}{r} = \pi d^2 \log \frac{R}{\varepsilon} \quad [15]$$

where we have used the Cauchy–Schwarz inequality for [14], and the characterization of the degree [7]. We may also observe that this lower bound is sharp if $\partial \varphi / \partial \tau$ is constant, that is, if the phase is $d\theta$ (and the vortex radial). The cost associated to $|u|$ in the energy imposes the length scale ε and is generally of

order 1 ($|\nabla u| \leq C/\varepsilon$), thus negligible compared to the cost associated to the phase, which blows up as $\log 1/\varepsilon$ as $\varepsilon \rightarrow 0$.

The above estimate is only valid as long as $B(x_0, R)$ does not contain any other zero of u . If vortices get close to each other or become numerous, one needs refined techniques to estimate their cost. This can be done through a “ball-construction method” introduced independently by Jerrard and Sandier.

Evaluating the Total Interaction Cost of Vortices

In a first approach, one studies configurations which satisfy the upper bound $E_{\varepsilon}(u) \leq C|\log \varepsilon|$. Then, lower bounds of the type [15] show that the total sum of the degrees (hence the total number of vortices of nonzero degree) remains bounded as $\varepsilon \rightarrow 0$. Up to extraction, we may assume these zeros a_i converge as $\varepsilon \rightarrow 0$ to a finite set of points p_i , with a total degree still denoted d_i . This can also be expressed as $\mu(u_{\varepsilon}) \rightarrow 2\pi \sum_i d_i \delta_{p_i}$ as $\varepsilon \rightarrow 0$.

This is not the only case of interest, since unbounded numbers of vortices do arise, especially in the physical situation of the energy with magnetic field, as we will see in the next section. However, this hypothesis, which was made in the work of Bethuel–Brezis–Hélein, makes the analysis easier and already allows us to exhibit the main phenomena.

Vortices in superconductors are generated by the presence of the external magnetic field h_{ex} . For the energy without magnetic field, this has to be replaced by some boundary condition which forces some degree. Bethuel–Brezis–Hélein considered the fixed Dirichlet boundary condition $u_{\varepsilon} = g$ on $\partial\Omega$, where g is a fixed unit-valued map on $\partial\Omega$, of degree $d > 0$. This forces u to have a total degree d in Ω . However, the Neumann boundary condition, for instance, can also be considered (the minimizers of E_{ε} are then simply constants, they are trivial, but one can still look for other critical points).

Let us return to lower bounds in order to look for the next order term in the energy (still with formal arguments). Cutting out holes $\cup_i B(p_i, \rho)$ of fixed size ρ around the limiting vortices p_i , we may assume that $u = e^{i\varphi}$ in $\Omega \setminus \cup_i B(p_i, \rho) = \Omega_{\rho}$, with φ a real-valued function, defined modulo 2π . Minimizing the energy outside of the holes amounts to solving

$$\begin{aligned} & \min_{\substack{u: \Omega_{\rho} \rightarrow S^1 \\ u=g \text{ on } \partial\Omega \\ \deg(u, \partial B(p_i, \rho))=d_i}} \frac{1}{2} \int_{\Omega_{\rho}} |\nabla u|^2 \end{aligned}$$

This is a harmonic map problem, whose solution is given in terms of φ by

$$\begin{aligned}\Delta\varphi &= 0 && \text{in } \Omega_\rho \\ \frac{\partial\varphi}{\partial\tau} &= \left\langle ig, \frac{\partial g}{\partial\tau} \right\rangle && \text{on } \partial\Omega \\ \int_{\partial B(p_i, \rho)} \frac{\partial\varphi}{\partial\tau} &= 2\pi d_i\end{aligned}$$

and in terms of the harmonic conjugate Φ which is the function (up to a constant) such that $\nabla\varphi = \nabla^\perp\Phi$,

$$\begin{aligned}\Delta\Phi &= 0 && \text{in } \Omega_\rho \\ \frac{\partial\Phi}{\partial\nu} &= \left\langle ig, \frac{\partial g}{\partial\tau} \right\rangle && \text{on } \partial\Omega \\ \int_{\partial B(p_i, \rho)} \frac{\partial\Phi}{\partial\nu} &= 2\pi d_i\end{aligned}\tag{16}$$

As $\rho \rightarrow 0$, Φ behaves like the solution of

$$\begin{aligned}\Delta\Phi_0 &= 2\pi \sum_i d_i \delta_{p_i} && \text{in } \Omega \\ \frac{\partial\Phi_0}{\partial\nu} &= \left\langle ig, \frac{\partial g}{\partial\tau} \right\rangle && \text{on } \partial\Omega\end{aligned}\tag{17}$$

Hence, we have

$$\begin{aligned}\frac{1}{2} \int_{\Omega_\rho} |\nabla\varphi|^2 &= \frac{1}{2} \int_{\Omega_\rho} |\nabla\Phi|^2 \\ &\simeq \frac{1}{2} \int_{\Omega_\rho} |\nabla\Phi_0|^2 \\ &= \pi \sum_i d_i^2 \log \frac{1}{\rho} + W_d(p_1, \dots, p_n) \\ &\quad + o(1) \quad \text{as } \rho \rightarrow 0\end{aligned}\tag{18}$$

where

$$\begin{aligned}W_d(a_1, \dots, a_n) &= -\pi \sum_{i \neq j} d_i d_j \log |p_i - p_j| \\ &\quad - \pi \sum_i d_i R(a_i) \\ &\quad + \frac{1}{2} \int_{\partial\Omega} \Phi_0 \left(ig, \frac{\partial g}{\partial\tau} \right)\end{aligned}\tag{19}$$

and $R(x) = \Phi_0(x) - \sum_i d_i \log |x - p_i|$. The function W was introduced by Bethuel–Brezis–Hélein and called the renormalized energy, since it consists in the part of the energy that is left after subtracting the “infinite part” in $|\log \varepsilon|$ from E_ε . It contains the (logarithmic) interaction energy between the vortices: we see that vortices with degrees of same sign repel one another while vortices with degrees of opposite signs attract one another. The $\pi d_i^2 \log 1/\rho$ term corresponds to the self-interaction, or cost of

the vortex of core of size ρ ; it is what replaces the infinite term in the formal calculation.

Now [18] is a good estimate for the optimal energy outside of the holes, while the energy in holes of size ρ can be bounded below by [15]. Given the degree d_i on the boundary $\partial B(p_i, \rho)$ of the small hole, $B(p_i, \rho)$ contains one or several zeros of u of degrees δ_k with total degree $\sum_k \delta_k = d_i$. In view of [15], since the cost of a vortex of degree d grows like $\pi d^2 |\log \varepsilon|$, and since the infimum of $\sum_k \delta_k^2$ under the constraint $\sum_k \delta_k = d_i$ is $\delta_k = \text{sign}(d_i)$, the least costly way to achieve this is to have $|d_i|$ vortices of degree $\text{sign}(d_i)$. The smallest lower bound possible is thus

$$\frac{1}{2} \int_{B(p_i, \rho)} |\nabla u|^2 + \frac{(1 - |u|^2)^2}{2\varepsilon^2} \geq \pi |d_i| \log \frac{\rho}{\varepsilon} + C\tag{20}$$

where the constant C can be described explicitly. Adding up the results of [20] and [18], we find

$$\begin{aligned}E_\varepsilon(u) &\geq \pi \sum_i d_i^2 \log \frac{1}{\rho} \\ &\quad + \pi \sum_i |d_i| \log \frac{\rho}{\varepsilon} + W_d(p_1, \dots, p_n) \\ &\quad + nC + o_\rho(1) + o_\varepsilon(1) \\ &\geq \pi \sum_i |d_i| \log \frac{1}{\varepsilon} + W_d(p_1, \dots, p_n) \\ &\quad + nC + o_\varepsilon(1)\end{aligned}\tag{21}$$

with equality only if u has $|d_i|$ zeros of degree $\text{sign}(d_i)$ in each $B(p_i, \rho)$.

This provides a lower bound of the energy in terms of the vortices. Moreover, this bound is sharp: one can construct test configurations which have the given limiting vortices (p_i, d_i) , and an energy equal to the right-hand side of [21].

One can thus deduce the behavior of global minimizers of the energy. Given the total degree $d = \deg(g) > 0$ on $\partial\Omega$, we need $\sum_i d_i = d$, and the lowest value achievable under this constraint in the right-hand side of [21] is to have $d_i = 1$ for every i , and thus to have exactly d vortices of degree 1. Moreover, the limiting points p_i 's should minimize W . We thus are led to the first main result.

Theorem 1 (Bethuel–Brezis–Hélein). *Minimizers of E_ε under the boundary condition $u = g$, $\deg(g) = d > 0$, have d zeros of degree 1, which converge as $\varepsilon \rightarrow 0$ to a minimizer of W .*

This result can be rephrased as a result of Γ -convergence of $E_\varepsilon - \pi d |\log \varepsilon|$. It reduces the minimization of E_ε to one of W , which is a finite-dimensional problem (interaction of point charges).

Thus, we see again the interest of studying this asymptotic limit $\varepsilon \rightarrow 0$ because the vortices become pointlike and the problem reduces to a finite-dimensional one, or one of minimizing the vortex interaction.

Further Results

A nonvariational approach also allowed Bethuel–Brezis–Hélein to prove a further correspondence between E_ε and W : they obtained that critical points of E_ε , under the upper bound $E_\varepsilon \leq C|\log \varepsilon|$, have vortices which converge to a critical point of W . Other important results are the study of the blow-up profiles or solutions in the whole plane, by Brezis–Merle–Rivière and Mironescu.

In two dimensions, the variational approach is also used to solve inverse problems (construct solutions) and study variants of the energy with pinning (or weighted) terms.

The variational approach is also fruitful in higher dimensions. In dimension 3, for example, vortices are not points but vortex lines, and the Jacobian $Ju = d(iu, du)$ can be seen as a current carried by the vortex line, with $\|Ju\|$ total mass of the current equal to π times the length of the line, and it was established by Jerrard and Soner that Ju_ε is compact in some weak sense, and converges, up to extraction, to some π times integer-multiplicity rectifiable current J , with

$$\liminf_{\varepsilon \rightarrow 0} \frac{E_\varepsilon(u_\varepsilon)}{|\log \varepsilon|} \geq \|J\|$$

In fact, a complete Γ -convergence result of $E_\varepsilon/|\log \varepsilon|$ can be proved, see the work of Alberti–Baldo–Orlandi, and thus minimizing E_ε reduces at the limit to minimizing the length of the line, leading to straight lines, or in higher dimensions, to codimension-2 minimal currents. This is a nontrivial problem, contrarily to dimension 2, where the Γ -limit of $E_\varepsilon/|\log \varepsilon|$ is trivial, which required to go to the lower-order term to find the nontrivial renormalized energy limit W .

The Functional with Magnetic Field

The aim here is to achieve the same objective: express or bound from below the energy by terms which depend only on the vortices and their degrees. The method consists in transposing the type of analysis above taking into account the magnetic field contribution to see how the external field triggers the sudden appearance of vortices, and for what values they appear (thus retrieving the critical fields, etc.). One of the main difficulties consists in the fact that the number of vortices becomes divergent,

which requires more delicate estimates. Also, it is then no longer possible to study the convergence of the individual zeros of ψ , so one studies instead the limit of rescalings of the vorticity measures $\mu(\psi, A)$.

Splitting of the Energy and Main Results

Let us recall that in the case with magnetic field, the vorticity is given by [12]. In addition, we may assume that the second set of equations in [5]

$$-\nabla^\perp h = j \text{ in } \Omega, \quad h = h_{\text{ex}} \text{ on } \partial\Omega \tag{22}$$

is satisfied (if not, keeping ψ fixed and choosing A which satisfies this equation always decreases the energy). Taking the curl of this equation, we find exactly

$$\begin{aligned} -\Delta h + h &= \mu(\psi, A) && \text{in } \Omega \\ h &= h_{\text{ex}} && \text{on } \partial\Omega \end{aligned} \tag{23}$$

Thus, the vorticity and the induced magnetic field are in one-to-one correspondence with each other. Combining it to the relation [13], we are led to the approximate relation

$$\begin{aligned} -\Delta h + h &\simeq 2\pi \sum_i d_i \delta_{a_i} && \text{in } \Omega \\ h &= h_{\text{ex}} && \text{on } \partial\Omega \end{aligned} \tag{24}$$

where again the a_i ’s are the vortex centers and d_i ’s their degrees, well known in physics as the “London equation.” It shows how the magnetic field is induced by the vortices which act like “charges,” and how the magnetic field “penetrates the sample” around the positive vortex locations. Of course this equation is only an approximation, because the singularities at the a_i ’s, where h would become infinite, are really smoothed out in $\mu(\psi, A)$; however, the approximation is good far from the vortex cores, just as [17] is an approximation for [16].

It is then natural to introduce the field corresponding to the vortex-free situation, which is $h_{\text{ex}} h_0$ where h_0 solves

$$\begin{aligned} -\Delta h_0 + h_0 &= 0 && \text{in } \Omega \\ h_0 &= 1 && \text{on } \partial\Omega \end{aligned} \tag{25}$$

h_0 is thus a fixed smooth function, depending only on Ω , and when there are no vortices, we expect h to be approximately $h_{\text{ex}} h_0$. Moreover, $h' := h - h_{\text{ex}} h_0$ then solves

$$\begin{aligned} -\Delta h' + h' &= \mu(\psi, A) \simeq 2\pi \sum_i d_i \delta_{a_i} && \text{in } \Omega \\ h' &= 0 && \text{on } \partial\Omega \end{aligned} \tag{26}$$

Defining the Green kernel $G(\cdot, y)$ by

$$\begin{aligned} -\Delta G + G &= \delta_y & \text{in } \Omega \\ G &= 0 & \text{on } \partial\Omega \end{aligned} \quad [27]$$

and S by $S(x, y) = 2\pi G(x, y) + \log|x - y|$, for x far enough from the a_i 's, we may approximate h' by

$$h'(x) = 2\pi \sum_i G(x, a_i) \quad [28]$$

Using the second Ginzburg–Landau equation [22] and the fact that $|\psi| \leq 1$, we have $|\nabla_A \psi| \geq |j| = |\nabla h|$, thus $G_\varepsilon(\psi, A) \geq (1/2) \int_\Omega |\nabla h|^2 + |h - h_{\text{ex}}|^2$. Plugging in the decomposition $h = h_{\text{ex}} h_0 + h'$ and using an integration by parts and [26], one finds

$$\begin{aligned} G_\varepsilon(\psi, A) &= \frac{1}{2} h_{\text{ex}}^2 \int_\Omega |\nabla h_0|^2 + |h_0 - 1|^2 \\ &\quad + h_{\text{ex}} \int_\Omega \nabla h_0 \cdot \nabla h' + (h_0 - 1) h' \\ &\quad + \frac{1}{2} \int_\Omega |\nabla h'|^2 + |h'|^2 \\ &= h_{\text{ex}}^2 J_0 + h_{\text{ex}} \int_\Omega (h_0 - 1) \mu(\psi, A) \\ &\quad + \frac{1}{2} \int_\Omega |\nabla h'|^2 + |h'|^2 \end{aligned} \quad [29]$$

where J_0 is the constant $(1/2) \int_\Omega |\nabla h_0|^2 + |h_0 - 1|^2$. The right-hand side of eqn [29] can be expressed in terms of the vortices. First, using [26], we have $\int_\Omega (h_0 - 1) \mu(\psi, A) \simeq 2\pi \sum_i d_i (h_0 - 1)(a_i)$. Second, the expression $\int_\Omega |\nabla h'|^2 + |h'|^2$ can be treated exactly like $E_\varepsilon(u)$ in the previous section, using lower bounds for the cost of vortices provided by the Jerrard–Sandier method, we are led to the (approximate) relation

$$\begin{aligned} \frac{1}{2} \int_\Omega |\nabla h'|^2 + |h'|^2 &\geq \pi \sum_i |d_i| \log \frac{1}{\varepsilon} \\ &\quad - \pi \sum_{i \neq j} d_i d_j \log |a_i - a_j| \\ &\quad + \pi \sum_{i,j} d_i d_j S(a_i, a_j) \end{aligned} \quad [30]$$

Combining this to [29] we find the decomposition

$$\begin{aligned} G_\varepsilon(\psi, A) &\geq h_{\text{ex}}^2 J_0 + \pi \sum_i |d_i| \log \varepsilon \\ &\quad + 2\pi h_{\text{ex}} \sum_i d_i (h_0 - 1)(a_i) \\ &\quad - \pi \sum_{i \neq j} d_i d_j \log |a_i - a_j| \\ &\quad + \pi \sum_{i,j} d_i d_j S(a_i, a_j) \end{aligned} \quad [31]$$

On the other hand, this inequality is sharp: as before, given vortices a_i , one can construct a

configuration (ψ, A) for which this is an equality, at leading order.

In that relation, $h_{\text{ex}}^2 J_0$ is a fixed energy, the energy of the vortex-free configuration. To it are added the intrinsic cost of each vortex $\pi |d_i| |\log \varepsilon|$, the interaction cost between vortices, and the interaction between the vortices and the external field $2\pi h_{\text{ex}} \sum_i d_i (h_0 - 1)(a_i)$.

It is then simple, by minimizing the right-hand side with respect to the vortices for a given h_{ex} , and observing that $h_0 - 1 \leq 0$, to deduce a few basic facts about vortices: vortices of positive degree (and of degree $+1$) are preferred, each vortex costs $\pi |\log \varepsilon|$, and allows to gain at best an energy $2\pi h_{\text{ex}} \max |h_0 - 1|$ when placed at the minimum of $h_0 - 1$. Therefore, vortices become favorable when their cost becomes smaller than the gain, that is, when h_{ex} becomes larger than the “first critical field”

$$H_{c_1} \sim \frac{|\log \varepsilon|}{2|\min(h_0 - 1)|} \quad [32]$$

We have the first main result.

Theorem 2 (Sandier–Serfaty). *When ε is small enough and $h_{\text{ex}} \leq H_{c_1}$, then minimizers of G_ε have no vortices.*

On the other hand, if $h_{\text{ex}} \geq H_{c_1}$, the vortices cannot all be located at the same minimum point of $h_0 - 1$, because their repulsion $-\pi \sum_{i \neq j} \log |a_i - a_j|$ would be infinite. There is thus a trade-off between their repulsion and the cost for being far from the minimum of $h_0 - 1$. Only if n , the number of vortices, is small compared to h_{ex} do the vortices tend to concentrate near the minimum of $h_0 - 1$. If so, then, assuming for simplicity that the minimum of $h_0 - 1$ is achieved at a unique point p , and denoting by Q the Hessian of $h_0 - 1$ at p , in the relation above $(h_0 - 1)(a_i)$ can be approximated by $\min(h_0 - 1) + (1/2)Q(a_i - p)$ and thus $G_\varepsilon(\psi, A)$ by

$$\begin{aligned} G_\varepsilon(\psi, A) &\sim h_{\text{ex}}^2 J_0 + \pi n |\log \varepsilon| + 2\pi n h_{\text{ex}} \min(h_0 - 1) \\ &\quad + \pi h_{\text{ex}} \sum_i Q(a_i - p) \\ &\quad - \pi \sum_{i \neq j} d_i d_j \log |a_i - a_j| + \pi n^2 S(p, p) \end{aligned} \quad [33]$$

From this relation, optimizing on ℓ , the characteristic distance to p and characteristic distance between the vortices, we find that $\ell = \sqrt{n/h_{\text{ex}}}$ is optimal.

Moreover, optimizing with respect to n , we find that n should remain bounded (as $\varepsilon \rightarrow 0$) when $h_{\text{ex}} \leq H_{c_1} + O(|\log |\log \varepsilon||)$. In that regime, rescaling by setting $x_i = ((a_i - p)/\ell)$, we have the following result:

Theorem 3 (Sandier–Serfaty). *There exist fields $H_n \sim H_{c_1} + C(n-1) \log |\log \varepsilon|$ such that when $H_n \leq h_{\text{ex}} < H_{n+1}$, minimizers of G_ε have n vortices of degree 1, and the rescaled vortices x_i ’s tend to minimize:*

$$w_n(x_1, \dots, x_n) = -\pi \sum_{i \neq j} \log |x_i - x_j| + \pi n \sum_{i=1}^n Q(x_i) \quad [34]$$

If $h_{\text{ex}} - H_{c_1} \gg \log |\log \varepsilon|$, then the optimal number of vortices n becomes unbounded as $\varepsilon \rightarrow 0$. The analysis above still holds, but in order to get a convergence of the vortices, one needs to rescale the vorticity measure by n . There is an intermediate regime, for $\log |\log \varepsilon| \ll h_{\text{ex}} - H_{c_1} \ll |\log \varepsilon|$ for which n should be $\gg 1$ but still $n \ll h_{\text{ex}}$, so $\ell \ll 1$: vortices are numerous, but still concentrate around p . Rescaling by the scale ℓ as above, we prove that the density of vortices (after dividing it by n) converges to a probability measure, minimizer of the energy

$$I(\mu) = -\pi \int_{\mathbb{R}^2 \times \mathbb{R}^2} \log |x - y| d\mu(x) d\mu(y) + \pi \int_{\mathbb{R}^2} Q(x) d\mu(x) \quad [35]$$

This is an averaged/continuous form of [34].

If $h_{\text{ex}} - H_{c_1}$ is of order $|\log \varepsilon|$, then the optimal number n becomes of order h_{ex} and the vortices no longer concentrate around a single point.

The simplest approach is then to simply consider the vorticity measure $\mu(\psi, A)$ and to rescale it by the order n , hence by h_{ex} . Then $(1/h_{\text{ex}})\mu(\psi, A)$ converges, after extraction, to some measure μ_* . A continuous version of [31] can thus be written, using [12], as

$$G_\varepsilon(\psi, A) \geq \frac{1}{2} h_{\text{ex}} |\log \varepsilon| \int_{\Omega} |\mu_*| + \frac{1}{2} h_{\text{ex}}^2 \int_{\Omega} |\nabla h_{\mu_*}|^2 + |h_{\mu_*}|^2 \quad [36]$$

where h_{μ_*} solves

$$\begin{aligned} -\Delta h_{\mu_*} + h_{\mu_*} &= \mu_* & \text{in } \Omega \\ h_{\mu_*} &= 1 & \text{on } \partial\Omega \end{aligned}$$

Again, this inequality can be proved to be sharp (by a construction) and allows to show that minimizers of G_ε have a vorticity $\mu(\psi, A)$ such that $\mu(\psi, A)/h_{\text{ex}}$ converges to a minimizer of

$$\mathcal{G}(\mu_*) = \frac{1}{2} \left(\lim_{\varepsilon \rightarrow 0} \frac{|\log \varepsilon|}{h_{\text{ex}}} \right) \int_{\Omega} |\mu_*| + \frac{1}{2} \int_{\Omega} |\nabla h_{\mu_*}|^2 + |h_{\mu_*}|^2$$

In fact the stronger result holds, in that sense:

Theorem 4 (Sandier–Serfaty). $G_\varepsilon/h_{\text{ex}}^2$ Γ -converges to \mathcal{G} .

The limit problem of minimizing \mathcal{G} turns out to have a simple solution in terms of an obstacle problem: the optimal μ_* is a uniform density of vortices on a subdomain of Ω determined through a free boundary problem (and depending on h_{ex}), which is nonzero.

In all these regimes, we have thus been able to identify the optimal number and repartition of vortices through a Γ -convergence-type approach, that is, by reducing the minimization of the energy to the minimization of a limiting problem: w_n or I or \mathcal{G} , according to the regime.

Further Results

Concerning vortices, in the same spirit as what was done for E_ε , we can obtain necessary conditions characterizing limiting vorticities obtained from sequences of (nonminimizing) critical points of the energy G_ε . They consist in passing to the limit in the conservative form of the Ginzburg–Landau equations [5].

Most of the results concerning the phase transitions at the next critical fields H_{c_2} and H_{c_3} are also obtained by nonvariational methods, and often by linear analysis.

The study of the Ginzburg–Landau energy in non-simply-connected domains is also very interesting because it leads to nontrivial topological effects, since in such domains there exist unit-valued maps with nonzero degree (corresponding to permanent currents).

See also: Abelian Higgs Vortices; Aharonov–Bohm Effect; Bose–Einstein Condensates; Gamma-Convergence and Homogenization; Gauge Theory: Mathematical Applications; Ginzburg–Landau Equation; High T_c Superconductor Theory; Image Processing; Mathematics; Superfluids; Topological Defects and Their Homotopy Classification; Variational Techniques for Microstructures.

Further Reading

- Bethuel F, Brezis H, and Hélein F (1994) *Ginzburg–Landau Vortices*. Boston: Birkhäuser.
- DeGennes PG (1966) *Superconductivity of Metal and Alloys*. New York: Benjamin.
- Jerrard RL and Soner HM (2002) The Jacobian and the Ginzburg–Landau energy. *Calculus of Variations and Partial Differential Equations* 14(2): 151–191.
- Sandier E and Serfaty S *Vortices in the Magnetic Ginzburg–Landau Model*. Birkhäuser (monograph to appear).
- Tinkham M (1996) *Introduction to Superconductivity*, 2nd edn. McGraw-Hill.

Variational Techniques for Microstructures

G Dolzmann, University of Maryland, College Park, MD, USA

© 2006 Elsevier Ltd. All rights reserved.

Austenite–Martensite Transformations and the Shape Memory Effect

Microstructures in materials that typically form in response to phase transformations in the solid state, and their impact on the elastic properties of these materials have been known for centuries. The discovery of the complex phase diagram of iron revolutionized the production of steels at the end of the nineteenth century. Starting in the 1980s, the mathematical description of microstructures in the framework of nonlinear elasticity has led to deep analytical questions and surprising developments in the calculus of variations and in nonlinear partial differential equations.

The mathematical approach outlined here is based on the following fundamental assumptions:

1. The observed configurations correspond to minimizers of or elements of minimizing sequences for an energy functional.
2. The qualitative properties of low energy states are determined from the set of minima of the free energy density.

Under these assumptions one aims at explaining experimental observations and to predict material properties based on minimizing an energy functional of the form

$$I(u) = \int_{\Omega} W(Du) \, dx$$

Here Ω is an ideal, unstressed reference configuration in \mathbb{R}^n , $u: \Omega \rightarrow \mathbb{R}^m$ is an elastic deformation, and $W: \mathbb{M}^{m \times n} \rightarrow \mathbb{R}$ is the stored energy density. In the case of physical interest, $m = n = 2$ or $m = n = 3$. For applications in elasticity we assume that $m = n$, but this assumption is not needed in the general theory. The energy density W and its structure depend critically on the temperature. However, since we are interested in the analysis of the material at a given temperature, we do not include this dependence explicitly.

The key ingredient of this model is the stored energy density W which has to reflect the properties of the specific material one wants to model. Frequently these are alloys, in particular shape memory alloys that undergo an austenite–martensite transformation. For most materials a closed analytic

expression for W is not available. In the spirit of the fundamental assumption (2) one therefore focuses on the structure of the set of minima of W which is determined from general invariance and symmetry principles. We may assume that $W \geq 0$ and that $K = \{X: W(X) = 0\} \neq \emptyset$. The principle of material frame indifference then asserts that

$$W(RF) = W(F) \quad \text{for all } R \in \text{SO}(n)$$

Here $\text{SO}(n)$ is the group of proper rotations, that is, the set of all matrices $R \in \mathbb{M}^{n \times n}$ with $R^T R = \text{Id}$ and $\det R = 1$.

The symmetry of the austenitic (high-temperature) phase implies that the energy density in the martensitic (low-temperature) phase is invariant under all changes of basis that leave the underlying lattice in the austenitic phase invariant. Therefore,

$$W(R^T F R) = W(F) \quad \text{for all } R \in \mathcal{P}_a$$

where \mathcal{P}_a is the point group of the austenite. In the case of a cubic to tetragonal phase transformation, this leads to $K = \text{SO}(3)U_1 \cup \text{SO}(3)U_2 \cup \text{SO}(3)U_3$

$$[1]$$

with

$$U_i = \eta^2 e_i \otimes e_i + \frac{1}{\eta} (I - e_i \otimes e_i) \quad [2]$$

in the martensitic phase (see Figure 1). A set of the form $\text{SO}(n)U_i$ is often referred to as an energy well.

The origin of the shape memory effect is the availability of a rich class of geometric patterns in which the martensitic phases can be arranged, thus leading to a great flexibility of the material to accommodate macroscopic deformations. Upon heating of the material above the transformation temperature, the martensitic phases lose their stability and the material returns to its unique shape in the

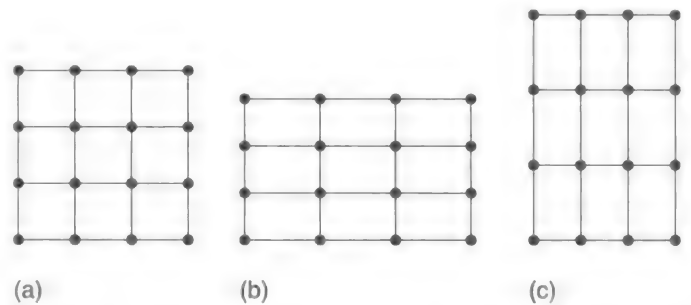


Figure 1 Two-dimensional cartoon of a cubic to tetragonal phase transformation in a single crystal: (a) a cubic lattice, (b) and (c) tetragonal variants which are stretched in directions e_1 and e_2 , respectively. (Sketch not to scale.)

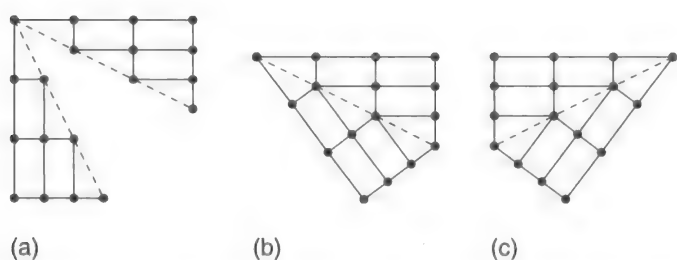


Figure 2 Formation of phase boundaries in a single crystal. (a) The upper right half of the lattice deforms into phase I with the constant deformation gradient U_1 , the lower left half of the lattice deforms into phase II with constant deformation gradient U_2 . (b) An additional rotation is needed to accomplish a continuous deformation, see formula [3]. (c) A different configuration with a different orientation of the interface. (Sketch not to scale.)

austenitic phase. The two solutions of Hadamard's compatibility condition

$$QU_2 - U_1 = a \otimes b, \quad Q \in \text{SO}(3)$$

are given by

$$Q_1 = \frac{1}{\eta^4 + 1} \begin{pmatrix} 2\eta^2 & \eta^4 - 1 & 0 \\ 1 - \eta^4 & 2\eta^2 & 0 \\ 0 & 0 & \eta^4 + 1 \end{pmatrix} \quad [3]$$

and $Q_2 = Q_1^T$ (see Figure 2). The normals (in the reference configuration) are given by $(1, \pm 1, 0)/\sqrt{2}$. It is one of the successes of the theory that it provides an analytical derivation of the normals to the twinning planes.

The Direct Method in the Calculus of Variations

The mathematical interest in the variational problems described in the previous section lies in the fact that existence of minimizers cannot in general be obtained by a straightforward application of the direct methods in the calculus of variations. This approach is based on the idea to (1) choose a minimizing sequence for the functional I , (2) show that this sequence is bounded and precompact, and (3) prove that the functional is lower semicontinuous with respect to the notion of convergence,

$$I(u) \leq \liminf_{j \rightarrow \infty} I(u_j) \quad \text{if } u_j \rightarrow u$$

The typical choice is to seek u_j in a suitable Sobolev space $W^{1,p}(\Omega; \mathbb{R}^m)$ with $1 < p \leq \infty$ which is related to growth and coercivity conditions for the energy density W ,

$$c_1|F|^p - c_2 \leq W(F) \leq c_3(|F|^p + 1) \quad \text{for all } F \in \mathbb{M}^{m \times n} \quad [4]$$

This leads to weak compactness in $W^{1,p}(\Omega; \mathbb{R}^m)$ (weak-* compactness in $W^{1,\infty}(\Omega; \mathbb{R}^m)$) and to the requirement of sequential weak lower semicontinuity of the functional,

$$I(u) \leq \liminf_{j \rightarrow \infty} I(u_j) \quad \text{if } u_j \rightarrow u \text{ in } W^{1,p}(\Omega; \mathbb{R}^m)$$

(sequential weak-* lower semicontinuity for $p = \infty$). Morrey's fundamental work establishes a link between convexity conditions for the energy density and lower semicontinuity of the variational integral: under suitable growth and coercivity conditions, sequential weak-* lower semicontinuity is equivalent to quasiconvexity of the integrand.

Definition 1 A function $W: \mathbb{M}^{m \times n} \rightarrow \mathbb{R}$ is said to be quasiconvex at F if

$$\int_{\Omega} W(F) dx \leq \int_{\Omega} W(F + D\phi) dx \quad \text{for all } \phi \in W_0^{1,\infty}(\Omega; \mathbb{R}^m)$$

and for all open and bounded domains $\Omega \subset \mathbb{R}^n$ with $\mathcal{L}^n(\partial\Omega) = 0$. It is said to be quasiconvex if it is quasiconvex at all F .

In the language of nonlinear elasticity, W is quasiconvex if affine functions are minimizers of the energy functional subject to their own boundary conditions. The direct method implies the following classical existence theorem.

Theorem 1 Suppose that $W: \mathbb{M}^{m \times n} \rightarrow \mathbb{R}$ is quasiconvex and satisfies the growth and coercivity condition [4]. Let $u_0 \in W^{1,p}(\Omega; \mathbb{R}^m)$. Then the variational problem: minimize $I(u)$ in

$$\mathcal{A} = \{u \in W^{1,p}(\Omega; \mathbb{R}^m) : u - u_0 \in W_0^{1,p}(\Omega; \mathbb{R}^m)\}$$

has a minimizer.

The remarkable fact is that the structure of the zero set of a typical energy W modeling a phase-transforming material in its low-temperature phase prevents W from being quasiconvex. In order to see this, let $Q \subset \mathbb{R}^3$ be a cube with two of its sides perpendicular to $b = (1, 1, 0)/\sqrt{2}$ and let h be the 1-periodic function with $h' = 0$ on $(0, \lambda)$ and $h' = 1$ on $(\lambda, 1)$ with $\lambda \in (0, 1)$. Define $v_j(x) = U_1 x + ab(jx \cdot b)/j$ and

$$u_j(x) = \min\{v_j(x), \text{dist}(x, \partial Q)\} \\ = \min\{U_1 x + ab(jx \cdot b)/j, \text{dist}(x, \partial Q)\}$$

where $\text{dist}(x, \partial Q) = \inf\{\|x - y\|_{\infty}, y \in \partial Q\}$. Then $u_j \rightarrow u$, $u(x) = Cx$ strongly in $L^{\infty}(Q; \mathbb{R}^3)$ and weakly-* in $W^{1,\infty}(Q; \mathbb{R}^3)$ with $C = \lambda U_1 + (1 - \lambda)Q_1 U_2 \notin K$ where K is the zero set of W , see the previous section.

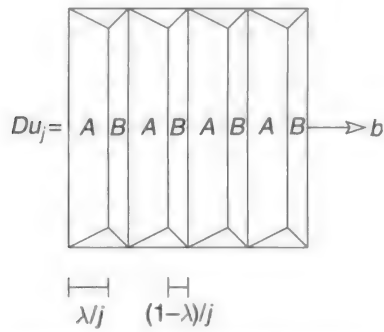


Figure 3 Construction of a minimizing sequence u_j with $Du_j \rightarrow \{A, B\}$ in measure and affine boundary conditions $u(x) = \lambda A + (1 - \lambda)B$. Hadamard's compatibility condition requires that $A - B = a \otimes b$ is a rank-1 matrix and that the planar interfaces are perpendicular to b .

Moreover, $Du_j \in \{U_1, Q_1 U_2\}$ except in a small transition layer of volume $\mathcal{O}(1/j)$ close to ∂Q and

$$I(u) = \int_{\Omega} W(C) dx > \liminf_{j \rightarrow \infty} I(u_j) = 0$$

This inequality shows that the functional is not weakly-* lower semicontinuous and therefore W fails to be quasiconvex. The oscillations of u_j on a scale $1/j$ are part of the mathematical model for the microstructures frequently observed in shape memory alloys. More generally, whenever u is a Sobolev function on a domain Ω such that Du takes only two values, say $Du \in \{A, B\}$, on open sets which are not empty and whose union is Ω (up to a set of measure zero), then the tangential continuity of the derivatives implies that the difference $A - B$ is a matrix of rank 1, $A - B = a \otimes b$, and that the interfaces between the regions with $Du = A$ and $Du = B$ are hyperplanes with normal parallel to b . This statement is usually referred to as “Hadamard’s compatibility condition.” Moreover, the pattern in **Figure 3** is known as a “simple laminate” and the matrices A and B are said to be rank-1 connected.

Relaxation

The discussion in the previous section shows that the variational problems related to models in materials science typically fail to be weakly lower semicontinuous. One approach which allows us to recover the macroscopic energy of the system and the macroscopic stress-strain relation is to pass to the relaxed variational problem which involves the quasiconvex envelope of the energy density W .

Definition 2 Let $W: \mathbb{M}^{m \times n} \rightarrow \mathbb{R}$ be given. The function

$$W^{qc} = \sup\{f: f \leq W, f \text{ quasiconvex}\}$$

is called the quasiconvex envelope of W . Equivalently,

$$W^{qc}(F) = \inf_{\phi \in W_0^{1,\infty}(\Omega; \mathbb{R}^m)} \frac{1}{|\Omega|} \int_{\Omega} W(F + D\phi) dx$$

This formula implies that W^{qc} is the macroscopic energy of the system in the sense that it characterizes the smallest energy per unit volume that is required to subject a volume element to a deformation with affine boundary conditions. Here the system is allowed to minimize its energy with microstructures at any scale, a mechanism which was already explored in the previous section. The arguments in this section prove that $W^{qc}(C) = 0$ and this shows that the zero set of W^{qc} can be strictly larger than the zero set of K , see Definition 4. The relaxed functional is given by

$$I^{qc}(u) = \int_{\Omega} W^{qc}(Du) dx$$

Since W^{qc} satisfies the growth and coercivity conditions [4] if they are satisfied by W , the functional I^{qc} attains its minimum subject to given boundary conditions. The functional I^{qc} is the weakly lower semicontinuous envelope of I in the sense that minimizing sequences for I contain subsequences that converge to minimizers of I^{qc} and for all u there exists a sequence u_j which converges in $W^{1,p}(\Omega; \mathbb{R}^m)$ to u such that the energies converge, $I(u_j) \rightarrow I(u)$. However, a lot of information in particular about oscillation patterns might be lost in the passage from I to I^{qc} since the knowledge of a minimizer u for I^{qc} does not provide any immediate information about the behavior of any minimizing sequence for I that converges to u . Moreover, the minimization problem required in the definition of the relaxed energy has been solved explicitly only for very special energy densities.

In this context, one often relies on two related notions of convexity, one sufficient and the other necessary for quasiconvexity. For $F \in \mathbb{M}^{m \times n}$ let $M(F) \in \mathbb{R}^{d(m,n)}$ be the vector of all minors (subdeterminants) of F . In the special case $m = n = 2$ we have $M(F) = (F, \det F) \in \mathbb{R}^5$ and for $m = n = 3$ we find $M(F) = (F, \text{cof } F, \det F) \in \mathbb{R}^{19}$ where $\text{cof } F$ is the 3×3 matrix of all 2×2 subdeterminants of F .

Definition 3 Let $W: \mathbb{M}^{m \times n} \rightarrow \mathbb{R}$ be given. The function W is said to be polyconvex if there exists a convex function $g: \mathbb{R}^{d(m,n)} \rightarrow \mathbb{R}$ such that $W(F) = g(M(F))$. The function W is rank-1 convex if it is convex along all rank-1 lines in $\mathbb{M}^{m \times n}$, that is, the function $t \mapsto W(F + tR)$ is convex for all $F \in \mathbb{M}^{m \times n}$ and all $R \in \mathbb{M}^{m \times n}$ with $\text{rank}(R) = 1$.

All notions of convexity reduce to classical convexity if $m=1$ or $n=1$. In the vector-valued case $m, n > 1$ the following implications are true:

$$\begin{aligned} f \text{ convex} &\Rightarrow f \text{ polyconvex} \Rightarrow f \text{ quasiconvex} \\ &\Rightarrow f \text{ rank-1 convex} \end{aligned}$$

The reverse statements for the first two implications are not true. Rank-1 convexity does not imply quasiconvexity for $m \geq 3$ and it is a fundamental open problem with deep connections to harmonic analysis to decide whether rank-1 convexity and quasiconvexity are equivalent for $m=n=2$.

The polyconvex and the rank-1 convex envelope of an energy density W are defined analogously to Definition 2. In view of the implications between the different notions of convexity, one has $W^{pc} \leq W^{qc} \leq W^{rc}$ and essentially all explicitly known relaxation formulas are based on the approach to construct a candidate W^* for W^{rc} and to verify that W^* is polyconvex. Then the inequalities become equalities and one obtains a characterization for the relaxed energy. This approach does not work for extended-valued functions which are used in models for incompressible materials since quasiconvexity does not imply rank-1 convexity in this case. However, for a model system of particular interest, nematic elastomers, a complete characterization of the relaxed energy, the macroscopic stress-strain relation, and the macroscopic phase diagram have been obtained.

Classical and Generalized Minimizers

The discussion of observed configurations as elements of minimizing sequences $\{u_j\}$ in the section “The direct method in the calculus of variations” leaves the question of the existence of minimizers open. The answer cannot be obtained via the direct methods since minimizing sequences do not need to converge strongly to minimizers. One approach to obtain the existence of solutions u with $I(u) = 0$ is to solve the differential relation $Du \in K, u(x) = Fx$ on $\partial\Omega$ by constructing special minimizing sequences that converge strongly so that one can pass to the limit in the energy integral. This idea has led to surprising solutions u with affine boundary conditions for the two-well problem where $K = SO(2)\text{diag}(\eta, 1/\eta) \cup SO(2)\text{diag}(1/\eta, \eta)$. However, the structure of the solutions is intrinsically complicated in the sense that the phase boundary has infinite length unless the boundary conditions are given by $u(x) = Fx$ with $F \in K$.

More generally, the right tool to pass to the limit in nonlinear functions of $z_j = Du_j$ like the energy is the

“Young measure” generated by a subsequence. It is given by a family of probability measures ν_x that provide statistical information about the distribution of the values of z_j close to a given point x . The existence and the fundamental properties of Young measures are described in the following theorem. For simplicity we assume that the sequence z_j is uniformly bounded.

Theorem 2 (Fundamental theorem on Young measures). *Let $E \subset \mathbb{R}^n$ be measurable, $\mathcal{L}^n(E) < \infty$, and let $z_j: E \rightarrow \mathbb{R}^d$ be a measurable and bounded sequence. Then there exists a subsequence z_k and a weakly- $*$ measurable map $\nu: E \rightarrow \mathcal{M}(\mathbb{R}^d)$ such that the following assertions are true:*

- (i) *The measures ν_x are non-negative probability measures.*
- (ii) *If there exists a compact set K such that $u_k \rightarrow K$ in measure, then $\text{supp } \nu_x \subset K$ for a.e. $x \in E$.*
- (iii) *If $f \in C(\mathbb{R}^d)$ and if $f(z_k)$ is relatively weakly compact in $L^1(E)$, then $f(z_k) \rightharpoonup \bar{f}$ in $L^1(E)$ where $\bar{f}(x) = \langle \nu_x, f \rangle$.*

Here $\langle \nu_x, f \rangle$ denotes the integration of the function f with respect to the measure ν_x . For example, the Young measure generated by the sequence Du_j constructed in the section “The direct method in the calculus of variations” generates the Young measure $\nu_x = (1/2)\delta_A + (1/2)\delta_B$ (see Figure 3) and

$$\begin{aligned} I(u_j) &= \int_{\Omega} W(Du_j) \, dx \\ &\rightarrow \int_{\Omega} \int_{\mathbb{M}^{m \times n}} W(Y) \, d\nu_x(Y) \, dx = 0 \end{aligned}$$

A Young measure generated by a sequence of gradients is called a gradient Young measure (GYM). It is said to be homogeneous if $\nu_x = \nu$ is independent of x . We restrict our attention in the following to homogeneous GYM’s generated by sequences that are bounded in L^∞ . The importance of quasiconvexity is also reflected in the following characterization of homogeneous GYM’s.

Theorem 3 *A non-negative probability measure ν is a GYM if and only if there exists a compact set $K \subset \mathbb{M}^{m \times n}$ with $\text{supp } \nu \subset K$ and Jensen’s inequality $\langle \nu, f \rangle \geq f(\langle \nu, \text{id} \rangle)$ holds for all quasiconvex functions $f: \mathbb{M}^{m \times n} \rightarrow \mathbb{R}$.*

This motivates to characterize the generalized limits of minimizing sequences as

$$\begin{aligned} \mathcal{M}^{qc}(K) &= \{ \nu \in \mathcal{M}(K): f(\langle \nu, \text{id} \rangle) \leq \langle \nu, f \rangle \\ &\text{for all } f: \mathbb{M}^{m \times n} \rightarrow \mathbb{R} \text{ quasiconvex} \} \end{aligned}$$

where $\mathcal{M}(K)$ is the set of all probability measures supported on K . If ν is generated by a sequence of

functions with affine boundary conditions $u_j(x) = Fx$, then $\langle \nu, \text{id} \rangle = F$. The set of all affine deformations of the material that can be recovered by heating (shape memory effect) is therefore given as the set of all centers of mass of homogeneous GYM's supported on K , the so-called "quasiconvex hull" K^{qc} of K .

Definition 4 Suppose that $K \subset \mathbb{M}^{m \times n}$ is compact. We define the quasiconvex hull of K by

$$K^{\text{qc}} = \{F = \langle \nu, \text{id} \rangle : \nu \in \mathcal{M}^{\text{qc}}(K)\}$$

There are several equivalent definitions of K^{qc} . The foregoing definition corresponds to the definition of the convex hull of a set as the set of all centers of mass of probability measures supported on K (which satisfy Jensen's inequality for all convex f). The set K^{qc} can also be defined as the set of all points that cannot be separated by quasiconvex functions from K or as the zero set of the quasiconvex envelope of the distance function to K . The "polyconvex hull" K^{pc} and the "rank-1 convex hull" K^{rc} are defined analogously by replacing quasiconvexity with polyconvexity and rank-1 convexity in the foregoing definitions. It follows that $K^{\text{rc}} \subset K^{\text{qc}} \subset K^{\text{pc}}$ and all of these inclusions can be strict.

A particularly useful set of conditions are the minors conditions

$$\langle \nu, M \rangle = M(\langle \nu, \text{id} \rangle)$$

for all minors M which follow from the weak continuity of the minors. For example, if $K = \{A, B\} \subset \mathbb{M}^{2 \times 2}$, then any probability measure supported on K is given by $\nu = \lambda \delta_A + (1 - \lambda) \delta_B$. The minors condition with $M = \det$ implies that

$$\begin{aligned} \det(\lambda A + (1 - \lambda)B) &= \det \langle \nu, \text{id} \rangle = \langle \nu, \det \rangle \\ &= \lambda \det A + (1 - \lambda) \det B \end{aligned}$$

This identity is equivalent to

$$\lambda(1 - \lambda) \det(A - B) = 0$$

and therefore the quasiconvex hull is equal to K if and only if $\det(A - B) \neq 0$. A very instructive example is the set $K = \{(1, 3), (-1, -3), (-3, 1), (3, -1)\}$ viewed as a subset of the space of all diagonal matrices in $\mathbb{M}^{2 \times 2}$. It is frequently referred to as a T_4 configuration. The rank-1 convex hull is equal to the quasiconvex hull and given by the four points, the line segments, and the square in the center, the polyconvex hull is bounded by four hyperbolic arcs, and the convex hull is the square with the points as corners, see Figure 4. It is remarkable that the rank-1 convex hull is strictly

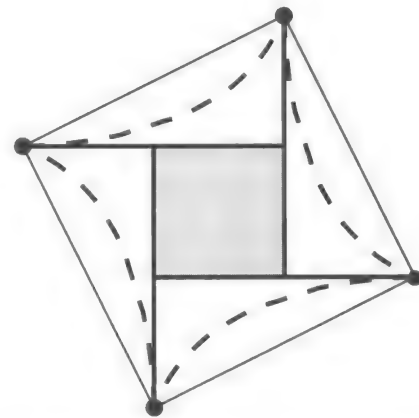


Figure 4 The four-point subset K in the space of all diagonal matrices and its convex hulls: $K^{\text{rc}} = K^{\text{qc}}$ are given by K , the line segments and the shade square, K^{pc} is bounded by the dashed hyperbolic arcs, and the convex hull is the outer square.

larger than the set K itself despite the fact that the set K does not contain any rank-1 connections.

There are only a few examples in which explicit characterizations of the convex hulls for sets invariant under $\text{SO}(n)$ have been obtained. For $K = \text{SO}(3)U_1 \cup \text{SO}(3)U_2$ (see [2]), one finds

$$K^{\text{qc}} = \left\{ F \in \mathbb{M}^{3 \times 3} : F^T F = \begin{pmatrix} a & c & 0 \\ c & b & 0 \\ 0 & 0 & 1/\eta^2 \end{pmatrix}, \right. \\ \left. ab - c^2 = \eta^2, a + b + 2|c| \leq \eta^4 + \frac{1}{\eta^2} \right\}$$

The quasiconvex hull of the three-well problem [1] is not known. In two dimensions one finds for

$$\begin{aligned} K &= \text{SO}(2)U_1 \cup \dots \cup \text{SO}(2)U_n, \\ \det U_i &= 1, i = 1, \dots, n \end{aligned}$$

that

$$K^{\text{qc}} = \left\{ F \in \mathbb{M}^{2 \times 2} : \det F = 1, |Fe|^2 \leq \max_{i=1, \dots, n} |U_i e|^2 \right\}$$

All examples in which envelopes of functions or hulls of sets have been obtained explicitly are based on the exceptional property that the polyconvex envelope coincides with the rank-1 convex envelope. The T_4 configuration in Figure 4 is one of the few cases where the quasiconvex hull is known to be different from the polyconvex hull. The construction of quasiconvex functions and the understanding of their properties is one of the challenges left for the future.

Bibliographical Remarks

This article can only review some of the highlights of mathematical developments related to models in

nonlinear elasticity for solid–solid phase transformations based on a huge body of work in the original literature. The precise references can be found in the extensive bibliographies of the books and review articles that are cited in the subsequent section, in particular in Ball (2004), Bhattacharya (2003), Dolzmann (2003), James and Hane (2000), and Müller (1999). This article focuses on models for single crystals; the behavior of polycrystals (which strongly depends on the amount of symmetry breaking in the transformation) was studied by Bhattacharya and Kohn.

The formulation of solid–solid phase transformations via nonlinear continuum theory goes back to Ericksen and the analysis via tools in the calculus of variations was initiated by Ball and James, Chipot and Kinderlehrer, and Fonseca. The Russian school developed the theory in linear elasticity in the 1960s, see Khachaturyan (1983) for a review. A detailed discussion of the crystallographic and group-theoretical aspects is contained in Pitteri and Zanzotto (2002).

Quasiconvexity was introduced by Morrey (1966) and his results were extended to Carathéodory integrands by Acerbi and Fusco and Marcellini. A modern treatment including Dacorogna's relaxation theorem and a summary of the various notions of convexity and their properties can be found in Dacorogna (1989). Šverák proved that rank-1 convexity does not imply quasiconvexity for $m \geq 3$ and Milton modified his example to show that the rank-1 convex hull of a set can be strictly smaller than its quasiconvex hull. The explicit characterizations for nematic elastomers were obtained by DeSimone and Dolzmann.

Lipschitz solutions to differential inclusions were constructed by Müller and Šverák based on Gromov's concept of convex integration, by Dacorogna and Marcellini using Baire's category argument, and by Kirchheim in the framework of Banach Mazur games. The structure of solutions of the two-well problem with finite surface energy was analyzed by Dolzmann and Müller. Young measures (also called parametrized measures or chattering controls) were originally introduced as generalized solutions for optimal control problems which do not admit classical solutions (Young 1969). Tartar (1979) introduced Young measures as a fundamental tool for the analysis of oscillation effects in partial

differential equations and for the passage from microscopic to macroscopic models. Gradient Young measures were characterized by Kinderlehrer and Pedregal. The four-point configuration was discovered independently in various contexts by several authors including Scheffer, Aumann and Hart, Casadio Tarabusi, Tartar, and Milton and Nesi. The characterization of the quasiconvex hull uses a quasiconvex function constructed by Šverák. The quasiconvex hull of the two-well problem in 3D was found by Ball and James, and the generalization to n wells in 2D by Bhattacharya and Dolzmann.

Acknowledgment

The work of G Dolzmann was supported by the NSF through grants DMS0405853 and DMS0104118.

See also: Gamma-Convergence and Homogenization; Variational Techniques for Ginzburg–Landau Energies.

Further Reading

- Ball JM (2004) Mathematical models of martensitic microstructure. *Materials Science and Engineering A* 378(1–2): 61–69.
- Bhattacharya K (2003) *Microstructure of Martensite*. Oxford: Oxford University Press.
- Dacorogna B (1989) *Direct Methods in the Calculus of Variations*. Berlin: Springer.
- Dolzmann G (2003) *Variational Methods for Crystalline Microstructure: Analysis and Computation*, Lecture Notes in Mathematics, vol. 1803. Berlin: Springer.
- James RD and Hane KF (2000) Martensitic transformations and shape-memory materials. *Acta Materiali* 48: 197–222.
- Khachaturyan A (1983) *Theory of Structural Transformations in Solids*. New York: Wiley.
- Morrey CB (1966) *Multiple Integrals in the Calculus of Variations*. Berlin: Springer.
- Müller S (1999) Variational methods for microstructure and phase transitions. *Proc. C.I.M.E. Summer School "Calculus of Variations and Geometric Evolution Problems," Cetraro, 1996*, Lecture Notes in Mathematics, vol. 1713. Berlin–Heidelberg: Springer.
- Pitteri M and Zanzotto G (2002) *Continuum Models for Phase Transitions and Twinning in Crystals*. London: Chapman and Hall.
- Tartar L (1979) Compensated compactness and partial differential equations. In: Knops R (ed.) *Nonlinear Analysis and Mechanics: Heriot–Watt Symposium*, vol. IV. London: Pitman.
- Young LC (1969) *Lectures on the Calculus of Variations and Optimal Control Theory*. Philadelphia–London–Toronto: Saunders.

Viscous Incompressible Fluids: Mathematical Theory

J G Heywood, University of British Columbia,
Vancouver, BC, Canada

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The Navier–Stokes equations

$$\rho(u_t + u \cdot \nabla u) = -\nabla p + \mu \Delta u + f \quad [1]$$

$$\nabla \cdot u = 0 \quad [2]$$

provide the simplest model for the motion of a viscous incompressible fluid that is consistent with the principles of mass and momentum conservation, and with Stokes' hypothesis that the internal forces due to viscosity must be invariant with respect to any superimposed rigid motion of the reference frame. Despite their simplicity, they seem to govern the motion of air, water, and many other fluids very accurately over a wide range of conditions. Thus, their mathematical theory is central to the rigorous analysis of many experimental observations, from the asymptotics of steady wakes and jets, to the dynamics of convection cells, vortex shedding, and turbulence. During the last 80 years, a great deal of progress has been made on both the basic mathematical theory of the equations and on its application to the understanding of such phenomena. But one of the most important matters, that of estimating the regularity of solutions over long periods of time, remains a vexing and fascinating challenge. Such an estimate will almost certainly be needed to prove the “global” existence of smooth solutions. By that we mean the existence of smooth solutions of the initial-value problem over indefinitely long periods of time without any restriction on the “size” of the data. To date we can prove the “local” existence of smooth solutions, but there remains a concern that if the data are large, solutions may develop singularities within a finite period of time. In fact, there is a great deal more at issue than this question of existence. A regularity estimate is required to prove the reliability of the equations as a predictive model. That is because any estimate for the continuous dependence of solutions on the prescribed data for a problem depends upon a regularity estimate, as do error estimates for numerical approximations. A global estimate for the regularity of solutions is also required for a mathematically rigorous theory of turbulence. In fact, it may be hoped that the insight which ultimately yields a global regularity estimate will

also be pivotal to our understanding of turbulence, perhaps justifying Kolmogoroff theory; see Heywood (2003). In this article we aim to present a relatively simple approach to the local existence, uniqueness, and regularity theory for the initial boundary value problem for the Navier–Stokes equations, and to discuss some observations that bear on the question of global regularity. A wider-ranging review of open problems is given in Heywood (1990), and further observations concerning the problem of global regularity are given in Heywood (1994).

Setting the Problem

To focus on core issues, we shall make some simplifying assumptions. The fluid under consideration will be assumed to completely fill (without free boundaries or vacuums) a bounded, connected, time-independent domain $\Omega \subset \mathbb{R}^n$, $n=2$ or 3 , with smooth boundary $\partial\Omega$. We are mainly interested in the three-dimensional case, but comparisons with the two-dimensional case are illuminating. The \mathbb{R}^n -valued velocity $u(x, t) = (u_1(x, t), \dots, u_n(x, t))$ and \mathbb{R} -valued pressure $p(x, t)$ are functions of the position $x = (x_1, \dots, x_n) \in \Omega$ and time $t \geq 0$. Equation [1] is an expression of Newton's second law of motion, equating mass density times acceleration on the left with several force densities on the right, due to pressure and viscosity, and sometimes a prescribed external force f . Written in full, using the summation convention over repeated indices, its i th component is

$$\rho \left(\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} \right) = -\frac{\partial p}{\partial x_i} + \mu \frac{\partial^2 u_i}{\partial x_j^2} + f_i$$

We will assume the density ρ and the coefficient of viscosity μ are positive constants.

In this article, we consider the initial boundary value problem consisting of the equations [1], [2] together with the initial and boundary conditions

$$u|_{t=0} = u_0, \quad u|_{\partial\Omega} = 0 \quad [3]$$

The initial velocity $u_0(x)$ is prescribed. It will be assumed to possess whatever smoothness is convenient, and to satisfy $\nabla \cdot u_0 = 0$ and $u_0|_{\partial\Omega} = 0$. The boundary condition is a reasonable one, since fluids adhere to rigid surfaces.

Notice that a further condition would be needed to uniquely determine the pressure, since only its derivatives appear in the problem as posed. We prefer to do without auxiliary conditions for the pressure, and to refer to u by itself as a solution of

the problem provided there exists a scalar function p which together with u satisfies [1]–[3]. The problem is said to be uniquely solvable if there is a unique solution u , in which case the gradient of the pressure is also uniquely determined, along with the pressure up to a constant. Notice also that under our assumptions a potential force like gravity has no effect on u . If u solves the problem in the absence of such a force, then the inclusion of the force affects only the pressure, from which the potential must be subtracted. It turns out that the inclusion of a prescribed nonpotential force, while complicating many of the estimates below, does not affect in any essential way those parts of the theory to be presented here. Thus, for simplicity, we shall henceforth assume that $f \equiv 0$.

Reynolds Number

We can make a slight further simplification of eqn [1] by rescaling, with the objective of setting $\rho = 1$, or even $\rho = 1$ and $\mu = 1$. This scaling is not required for the existence theory we are presenting, but provides an important insight for the study of stability, bifurcation, and turbulence. The Reynolds number

$$R = \frac{\max |u| \cdot |\Omega| \cdot \rho}{\mu}$$

plays an important role in rescaling. It expresses the ratio of the inertial to viscous effects. The notation $|\Omega|$ represents a characteristic length, such as the minimum diameter of a bounded domain. Generally speaking, a high Reynolds number corresponds to what is meant by “large” data, and the higher the Reynolds number the more inclined a flow is to instability and turbulence, and perhaps to the development of singularities. However, the size of the Reynolds number has precise implications only in comparing “dynamically similar” flows. We say that two vector fields $v(x, t)$ and $u(x, t)$ are dynamically similar if and only if $v(x, t) = \alpha u(x/\beta, t/\gamma)$ for some $\alpha, \beta, \gamma > 0$. In such a case, if u is defined in $\Omega \times [0, T)$, then v will be defined in $\beta\Omega \times [0, \gamma T)$, where $\beta\Omega = \{\beta x: x \in \Omega\}$. Furthermore, if u satisfies the Navier–Stokes equations, then v will satisfy

$$\begin{aligned} \rho\alpha^{-1}\gamma v_t + \rho\alpha^{-2}\beta v \cdot \nabla v \\ = -\beta\nabla p(x/\beta, t/\gamma) + \alpha^{-1}\beta^2\mu\Delta v \end{aligned} \quad [4]$$

which has the form of the Navier–Stokes equations if and only if the coefficients of the two inertial terms on the left-hand side are equal. That is, if and only if

$$\alpha\gamma = \beta \quad [5]$$

in which case

$$v_t + v \cdot \nabla v = -\nabla q + \eta\Delta v \quad [6]$$

with

$$\eta = \alpha\beta\mu/\rho \quad [7]$$

and $q(x, t) = \alpha^2\rho^{-1}p(x/\beta, t/\gamma)$. We refer to such u and v as dynamically similar flows. The relation [7], that follows from [5], is equivalent to the equality of the Reynolds numbers for the two flows,

$$\begin{aligned} R(u) &= \frac{\max |u| \cdot |\Omega| \cdot \rho}{\mu} \\ &= \frac{\max |\alpha u| \cdot |\beta\Omega| \cdot 1}{\eta} = R(v) \end{aligned}$$

The condition [5] can be satisfied simultaneously with the condition $\eta = 1$. For example, one may choose $\beta = 1$, $\alpha = \rho/\mu$, and $\gamma = \mu/\rho$. This achieves a rescaling of the equation to

$$v_t + v \cdot \nabla v = -\nabla q + \Delta v \quad [8]$$

without changing the domain. Different Reynolds numbers result from varying the magnitude of the velocity. In what follows, we will work with the Navier–Stokes equation in this simplest possible form.

Continuous Dependence on the Data

We begin our investigation of the initial boundary value problem

$$\begin{aligned} u_t + u \cdot \nabla u &= -\nabla p + \Delta u, \quad \nabla \cdot u = 0 \\ \text{for } (x, t) &\in \Omega \times (0, \infty), \\ u|_{t=0} &= u_0, \quad u|_{\partial\Omega} = 0 \end{aligned} \quad [9]$$

by considering two smooth solutions, say u and v , taking possibly different initial values u_0 and v_0 . Let their difference be $w = v - u$, with initial value w_0 , and let q be the difference of the corresponding pressures. Then, subtracting one equation from the other, one obtains

$$w_t + w \cdot \nabla w + u \cdot \nabla w + w \cdot \nabla u = -\nabla q + \Delta w \quad [10]$$

Multiplying this by w , integrating over Ω , and integrating by parts, one then obtains

$$\frac{1}{2} \frac{d}{dt} \|w\|^2 + \|\nabla w\|^2 = -(w \cdot \nabla u, w) \quad [11]$$

where

$$\begin{aligned}\|w\|^2 &= \int_{\Omega} w^2 \, dx \\ \|\nabla w\|^2 &= \int_{\Omega} \frac{\partial w_i}{\partial x_j} \frac{\partial w_i}{\partial x_j} \, dx \\ (w \cdot \nabla u, w) &= \int_{\Omega} w_j \frac{\partial u_i}{\partial x_j} w_i \, dx\end{aligned}$$

since (and this should further explain our notation)

$$\begin{aligned}(w_t, w) &= \int_{\Omega} w_t \cdot w \, dx \\ &= \frac{1}{2} \frac{d}{dt} \int_{\Omega} w^2 \, dx = \frac{1}{2} \frac{d}{dt} \|w\|^2 \\ (\Delta w, w) &= \int_{\Omega} \frac{\partial^2 w_i}{\partial x_j^2} w_i \, dx \\ &= - \int_{\Omega} \frac{\partial w_i}{\partial x_j} \frac{\partial w_i}{\partial x_j} \, dx = -\|\nabla w\|^2 \\ (\nabla q, w) &= \int_{\Omega} \frac{\partial q}{\partial x_i} w_i \, dx = - \int_{\Omega} q \frac{\partial w_i}{\partial x_i} \, dx = 0 \\ (u \cdot \nabla w, w) &= \int_{\Omega} u_j \frac{\partial w_i}{\partial x_j} w_i \, dx \\ &= - \frac{1}{2} \int_{\Omega} \frac{\partial u_i}{\partial x_j} w_i w_j \, dx = 0\end{aligned}$$

and similarly $(w \cdot \nabla w, w) = 0$. In deriving these we have used the fact that the vector fields are divergence free and vanish on the boundary. In the following, we will use such identities without further mention.

We can estimate the nonlinear term on the right-hand side of [11] by using the “Sobolev inequalities”

$$\begin{aligned}\|\phi\|_4^2 &\leq \|\phi\| \|\nabla \phi\|, & \text{if } n = 2 \\ \|\phi\|_4^2 &\leq \|\phi\|^{1/2} \|\nabla \phi\|^{3/2}, & \text{if } n = 3\end{aligned} \quad [12]$$

proved by Ladyzhenskaya (1969), though with larger constants. These are valid for any smooth function ϕ which vanishes on the boundary of Ω . It may be either scalar or vector valued. The norms on the left are L^4 -norms; we use the notation $\|\phi\|_p = (\int_{\Omega} |\phi|^p \, dx)^{1/p}$ for any $p > 1$, but usually drop the subscript when $p = 2$. Using first Hölder’s inequality and then [12], one obtains

$$\begin{aligned}|(w \cdot \nabla u, w)| &\leq \|w\|_4^2 \|\nabla u\| \\ &\leq \begin{cases} \|w\| \|\nabla w\| \|\nabla u\| & \text{if } n = 2 \\ \|w\|^{1/2} \|\nabla w\|^{3/2} \|\nabla u\| & \text{if } n = 3 \end{cases}\end{aligned}$$

Young’s inequality

$$ab \leq \frac{1}{p} a^p + \frac{1}{q} b^q$$

holds if $a, b > 0, p, q > 1$ and $1/p + 1/q = 1$. Taking $a = \sqrt{2} \|\nabla w\|$, along with $p = q = 2$ in the two-dimensional case, and $a = (4/3)^{3/4} \|\nabla w\|$, along with $p = 4/3, q = 4$ in the three-dimensional case, one obtains

$$\begin{aligned}|(w \cdot \nabla u, w)| &\leq \begin{cases} \|\nabla w\|^2 + \frac{1}{4} \|\nabla u\|^2 \|w\|^2, & \text{if } n = 2 \\ \|\nabla w\|^2 + \frac{27}{256} \|\nabla u\|^4 \|w\|^2, & \text{if } n = 3 \end{cases} \quad [13]\end{aligned}$$

Using these estimates for the right-hand side of [11], we obtain linear differential inequalities for $\|w\|^2$ that are easily integrated to give

$$\begin{aligned}\|w(t)\|^2 &\leq \begin{cases} \|w_0\|^2 \exp \int_0^t \frac{1}{2} \|\nabla u\|^2 \, d\tau, & \text{if } n = 2 \\ \|w_0\|^2 \exp \int_0^t \frac{27}{128} \|\nabla u\|^4 \, d\tau, & \text{if } n = 3 \end{cases} \quad [14]\end{aligned}$$

It follows that if we can estimate the integrals on the right, which concern only the solution u , and if v is a second solution, perhaps differing only slightly from u when $t = 0$, then we can estimate the difference $\|v(t) - u(t)\|$ at later times. Moreover, at any particular time this difference will be bounded proportionally to $\|v(0) - u(0)\|$. The integral on the right-hand side of the two-dimensional version of [14] is easily estimated using the energy estimate [16] below. The estimation of the corresponding integral in the three-dimensional case, without a restriction on the size of the data, remains an open problem. It can be regarded as the most important open problem in the Navier–Stokes theory. It would never be enough to somehow prove that solutions are smooth without estimating this integral, or something equivalent to it. Of course, if solutions were known to be smooth one could infer their uniqueness from [14], since smoothness would imply that the integrals are finite, which is enough to conclude that $\|w(t)\|$ is zero if $\|w_0\|$ is zero.

Energy Estimate

If one multiplies the Navier–Stokes equation for u by u , and proceeds as in deriving [11], one obtains

$$\frac{1}{2} \frac{d}{dt} \|u\|^2 + \|\nabla u\|^2 = 0 \quad [15]$$

and hence

$$\frac{1}{2} \|u(t)\|^2 + \int_0^t \|\nabla u\|^2 \, d\tau = \frac{1}{2} \|u_0\|^2 \quad [16]$$

This settles the matter of continuous dependence in the two-dimensional case. Together with [16], the two-dimensional version of [14] implies

$$\|w(t)\|^2 \leq \|w_0\|^2 \exp \frac{1}{4} \|u_0\|^2, \quad \text{if } n = 2 \quad [17]$$

We remark that the local rate of energy dissipation is $2|Du|^2$ rather than $|\nabla u|^2$, where Du is the stress tensor $Du = (1/2)(\nabla u + (\nabla u)^T)$. However, integrating over the domain, and integrating by parts using the boundary condition $u|_{\partial\Omega} = 0$, one may verify that the rate of total energy dissipation $2\|Du\|^2$ equals $\|\nabla u\|^2$. For the purpose of this article, it is convenient to write the energy identity as [15].

Estimates for $\|\nabla u(t)\|$ Pointwise in Time

Of course, an estimate for $\|\nabla u(t)\|$ pointwise in time would imply an estimate for the integral of $\|\nabla u(t)\|^4$ on the right-hand side of [14]. We can prove such an estimate for at least a finite interval of time by an argument due to Prodi (1962). It requires, in preparation, some deep results concerning the regularity of solutions of the steady Stokes equations. These cannot be proved here, but we can briefly summarize what will be needed. Let

$L^2(\Omega)$ = space of vector fields ϕ , with finite L^2 -norms $\|\phi\|$,

$C_0^\infty(\Omega)$ = space of smooth vector fields with compact support in Ω ,

$D(\Omega) = \{\phi \in C_0^\infty(\Omega) : \nabla \cdot \phi = 0\}$,

$J(\Omega)$ = completion of $D(\Omega)$ in the L^2 -norm $\|\phi\|$,

$J_1(\Omega)$ = completion of $D(\Omega)$ in the norm $\|\nabla \phi\|$,

$G(\Omega) = \{\nabla p : p \in L^2(\Omega) \text{ with } \nabla p \in L^2(\Omega)\}$, and

$P : L^2(\Omega) \rightarrow J(\Omega)$ be the L^2 -projection of $L^2(\Omega)$ onto $J(\Omega)$,

and define the Sobolev $W_2^2(\Omega)$ norm by

$$\begin{aligned} \|u\|_{W_2^2(\Omega)}^2 &= \|u\|^2 + \|\nabla u\|^2 \\ &\quad + \int_{\Omega} |\partial^2 u_i / \partial x_j \partial x_k|^2 dx \end{aligned}$$

Furthermore, observe that $(\nabla p, \phi) = 0$ for $\nabla p \in G(\Omega)$ and $\phi \in J(\Omega)$, since it holds if p is smooth and $\phi \in D(\Omega)$. Therefore, $P\nabla p = 0$, since $(P\nabla p, \phi) = (\nabla p, \phi) = 0$, for all $\phi \in J(\Omega)$. Later, when we need it, we will also argue that $L^2(\Omega) = J(\Omega) \oplus G(\Omega)$.

With these preparations, it is evident that every smooth vector field u satisfying $\nabla \cdot u = 0$ and $u|_{\partial\Omega} = 0$ can be regarded as a solution of the steady Stokes problem

$$-\Delta u + \nabla p = f \text{ and } \nabla \cdot u = 0 \text{ in } \Omega, \quad u|_{\partial\Omega} = 0 \quad [18]$$

with $f = -P\Delta u$. For such solutions, and hence for all such u , we have the estimates

$$\|u\|_{W_2^2(\Omega)} \leq c \|P\Delta u\| \quad [19]$$

and

$$\sup_{\Omega} |u|^2 \leq \begin{cases} c \|u\| \|P\Delta u\|, & \text{if } n = 2 \\ c \|\nabla u\| \|P\Delta u\|, & \text{if } n = 3 \end{cases} \quad [20]$$

with constants independent of u . It can also be shown that every such vector field u belongs to $J_1(\Omega)$ and hence to $J(\Omega)$; see Heywood (1973).

Some history and remarks are in order. The inequality [19] was proved independently by Solonnikov (1964, 1966), and by Prodi's student Cattabriga (1961). In fact, they gave L^p versions of it for all orders of the derivatives. Several proofs specific to the L^2 case needed here have been given by Solonnikov and Ščadilov (1973) and by Beirão da Veiga (1997). The inequalities [20] can be proved by combining [19] with appropriate Sobolev inequalities, or better, by combining [19] with recent inequalities of Xie (1991) which are of precisely the form [20], but with Δu instead of $P\Delta u$ on the right-hand side, and without the requirement that $\nabla \cdot u = 0$. The constant c in [19] depends upon the regularity of the boundary, and tends to infinity along with a bound for the boundary curvature. Through the work of Xie (1992, 1997), there is reason to believe that the inequalities [20] are probably valid for arbitrary domains, with the constant $c = (2\pi)^{-1}$ if $n = 2$, and $c = (3\pi)^{-1}$ if $n = 3$. Xie's efforts to prove this have been continued by the author (Heywood 2001). If the inequalities [20] can be proved for arbitrary domains (i.e., arbitrary open sets), with these fixed constants, then the approach to Navier–Stokes theory presented in this article will extend immediately to arbitrary domains, as explained in Heywood and Xie (1997), with estimates independent of the domain.

We go on now with an estimation of $\|\nabla u(t)\|$ based on [20]. Multiplying the Navier–Stokes equation for u by $-P\Delta u$, and integrating over Ω , one obtains

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|\nabla u\|^2 + \|P\Delta u\|^2 &= (u \cdot \nabla u, P\Delta u) \\ &\leq \sup_{\Omega} |u| \|\nabla u\| \|P\Delta u\| \quad [21] \end{aligned}$$

since $(u_t, -P\Delta u) = (Pu_t, -\Delta u) = (u_t, -\Delta u) = (\nabla u_t, \nabla u)$ and $(\nabla p, P\Delta u) = 0$.

The right-hand side of [21] can be estimated using [20] and Young's inequality:

$$\begin{aligned} & \sup_{\Omega} |u| \|\nabla u\| \|P\Delta u\| \\ & \leq \begin{cases} c\|u\|^{1/2} \|\nabla u\| \|P\Delta u\|^{3/2}, & \text{if } n = 2 \\ c\|\nabla u\|^{3/2} \|P\Delta u\|^{3/2}, & \text{if } n = 3 \end{cases} \\ & \leq \begin{cases} \frac{1}{2} \|P\Delta u\|^2 + c\|u\|^2 \|\nabla u\|^4, & \text{if } n = 2 \\ \frac{1}{2} \|P\Delta u\|^2 + c\|\nabla u\|^6, & \text{if } n = 3 \end{cases} \end{aligned}$$

Thus,

$$\begin{aligned} & \frac{d}{dt} \|\nabla u\|^2 + \|P\Delta u\|^2 \\ & \leq \begin{cases} c\|u\|^2 \|\nabla u\|^4, & \text{if } n = 2 \\ c\|\nabla u\|^6, & \text{if } n = 3 \end{cases} \end{aligned} \quad [22]$$

These differential inequalities are at the core of present theory. Consider first the two-dimensional case. It can be viewed as a linear differential inequality

$$\frac{d}{dt} \|\nabla u\|^2 \leq (c\|u\|^2 \|\nabla u\|^2) \|\nabla u\|^2 \quad [23]$$

with a coefficient $c\|u\|^2 \|\nabla u\|^2$ that is integrable, in view of the energy estimate [16]. Integrating it yields a "global" estimate; for all $t \geq 0$,

$$\begin{aligned} \|\nabla u(t)\|^2 & \leq \|\nabla u_0\|^2 \exp \int_0^t \|u\|^2 \|\nabla u\|^2 d\tau \\ & \leq \|\nabla u_0\|^2 \exp \frac{1}{2} \|u_0\|^4 \end{aligned} \quad [24]$$

However, if the three-dimensional version of [22] is viewed as a linear differential inequality, the coefficient to be integrated is $\|\nabla u(t)\|^4$. Thus, the same integral which is crucial to proving continuous dependence on the data is also crucial to proving regularity. What we can do in the three-dimensional case, is view [22] as a nonlinear differential inequality of the form

$$\varphi' \leq c\varphi^3 \quad \text{or} \quad \varphi' \leq c\|\nabla u\|^2 \varphi^2 \quad [25]$$

for $\varphi(t) = \|\nabla u(t)\|^2$. Integrating the first of these, one obtains a local estimate

$$\|\nabla u(t)\|^2 \leq \frac{\|\nabla u_0\|^2}{\sqrt{1 - 2c\|\nabla u_0\|^4 t}} \quad [26]$$

for

$$0 \leq t < \frac{1}{2c\|\nabla u_0\|^4}$$

without any restriction on the size of the data. Integrating the second, one obtains a global estimate

$$\begin{aligned} \|\nabla u(t)\|^2 & \leq \frac{\|\nabla u_0\|^2}{1 - c\|\nabla u_0\|^2 \int_0^t \|\nabla u\|^2 d\tau} \\ & \leq \frac{\|\nabla u_0\|^2}{1 - (c/2)\|u_0\|^2 \|\nabla u_0\|^2} \end{aligned} \quad [27]$$

valid for all $t \geq 0$, provided

$$\|u_0\|^2 \|\nabla u_0\|^2 < \frac{2}{c} \quad [28]$$

This is a good interpretation of what we mean by "small data." If Xie's conjecture is correct, that the constant in the three-dimensional version of [20] is $c = (3\pi)^{-1}$, then we obtain [25]–[28] with the constant $c = 3/(128\pi^2)$. Thus, $2/c \simeq 842$.

Further Regularity, Smoothing Estimates

Once one has an estimate of the form

$$\|\nabla u(t)\| \leq M(t), \quad \text{for } 0 \leq t < T \quad [29]$$

as provided by [24], [26], or [27], one can estimate the solution's derivatives of all orders over the open time interval $(0, T)$. The initial time $t = 0$ must be excluded from the interval, because the "imperfection" of prescribed data generally causes an impulsive acceleration along the boundary at time zero, resulting in a thin boundary layer in which the derivatives are so large that $\|\nabla u_t(t)\|$ and $\|u(t)\|_{W_2^3(\Omega)}$ tend to infinity as $t \rightarrow 0^+$. But the solution quickly smooths and remains smooth as long as [29] remains in force. Thus, our working assumption up to this point, that solutions are C^∞ smooth in $\bar{\Omega} \times [0, \infty)$ is not valid at $t = 0$. However, we will see that they are smooth in $\bar{\Omega} \times (0, T)$ and continuous in $\bar{\Omega} \times [0, T]$. They are also continuous on $[0, T]$ in the $W_2^2(\Omega)$ norm. This is sufficient regularity to justify everything that we have done to this point.

In this section, we give estimates for the derivatives of all orders with respect to time, of u and its first- and second-order derivatives with respect to space. In the next section, we will prove an existence theorem by Galerkin approximation. It will be easily seen that all of the estimates proved in this and previous sections, for solutions that are assumed to be smooth, also hold for the approximations, without any unproven assumptions. Therefore, they will be inherited by the solution that is obtained upon passing to the limit of the approximations. At first, this solution will be something of a generalized solution, not fully classical, but one which is C^∞ with respect to time over the interval $0 < t < T$, in the $W_2^2(\Omega)$ norm. In a final step,

viewing u at any fixed time as a solution of the steady Stokes equations, we can apply regularity estimates for the Stokes equations to infer that it is C^∞ in all variables throughout $\bar{\Omega} \times (0, T)$, with specific estimates for each derivative.

The estimates of this section are obtained by integrating an infinite sequence of differential inequalities, for $\|u\|, \|\nabla u\|, \|u_t\|, \|\nabla u_t\|, \|u_{tt}\|, \|\nabla u_{tt}\|, \dots$. The first two are [15] and [21], which have already been dealt with. It turns out that after these first two, each succeeding differential inequality is linearized by the estimates obtained from its predecessor, which explains why the time intervals for these additional estimates do not become successively shorter. In fact, in the two-dimensional case, the energy estimate resulting from [15], which is valid for all time, already gives the linearization [23] of [21], which then provides an estimate valid for all time. Except for noting such differences between the two- and three-dimensional cases, we will henceforth deal with only the three-dimensional case.

The differential inequalities just mentioned are obtained by estimating the right-hand sides of two sequences of differential identities, and ordering them by an iteration between the two sequences. The first sequence begins with and is patterned after the energy identity,

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|u\|^2 + \|\nabla u\|^2 &= 0 \\ \frac{1}{2} \frac{d}{dt} \|u_t\|^2 + \|\nabla u_t\|^2 &= -(u_t \cdot \nabla u, u_t) \\ \frac{1}{2} \frac{d}{dt} \|u_{tt}\|^2 + \|\nabla u_{tt}\|^2 &= -(u_{tt} \cdot \nabla u, u_{tt}) \\ &\quad - 2(u_t \cdot \nabla u_t, u_{tt}) \end{aligned} \quad [30]$$

etc.

while the second begins with and is patterned after Prodi's identity,

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|\nabla u\|^2 + \|P\Delta u\|^2 &= (u \cdot \nabla u, P\Delta u) \\ \frac{1}{2} \frac{d}{dt} \|\nabla u_t\|^2 + \|P\Delta u_t\|^2 &= (u_t \cdot \nabla u, P\Delta u_t) \\ &\quad + (u \cdot \nabla u_t, P\Delta u_t) \\ \frac{1}{2} \frac{d}{dt} \|\nabla u_{tt}\|^2 + \|P\Delta u_{tt}\|^2 &= (u_{tt} \cdot \nabla u, P\Delta u_{tt}) + \dots \end{aligned} \quad [31]$$

etc.

Before going on, notice that we can return to [22] and use [29] to infer a more complete estimate of the form

$$\begin{aligned} \|\nabla u(t)\|^2 + \int_0^t \|P\Delta u\|^2 d\tau \\ \leq B(M, t), \quad \text{for } 0 \leq t < T \end{aligned} \quad [32]$$

containing an integral of $\|P\Delta u\|^2$ on the left-hand side. We will use the notation $B(M, t)$ generically, for any bound that depends only on the function $M(t)$ and t . We remark, that a term $\|u_t\|^2$ can also be included under the integral sign on the left-hand side of [32], because $\|u_t\|$ and $\|P\Delta u\|$ are of essentially the same order, being the leading terms in the projection $u_t + P(u \cdot \nabla u) = P\Delta u$ of the Navier–Stokes equation. Finally, one can also include $\|u\|_{W_2^2(\Omega)}$ under the integral sign, in view of [19].

Going on, we obtain a third differential inequality from the second identity of the sequence [30]. Its right-hand side admits the estimate

$$\begin{aligned} -(u_t \cdot \nabla u, u_t) &\leq \|u_t\|_4^2 \|\nabla u\| \\ &\leq c \|u_t\|^{1/2} \|\nabla u_t\|^{3/2} \|\nabla u\| \\ &\leq \frac{1}{2} \|\nabla u_t\|^2 + c \|\nabla u\|^4 \|u_t\|^2 \end{aligned} \quad [33]$$

which, in view of [29] or [32], produces a linear differential inequality with integrable coefficients. Its integration yields an estimate of the form

$$\begin{aligned} \|u_t(t)\|^2 + \int_0^t \|\nabla u_t\|^2 d\tau \\ \leq B(M, t, \|u_t(0)\|), \quad \text{for } 0 \leq t < T \end{aligned} \quad [34]$$

provided $\|u_t(0)\|$ is bounded. Since $u_t = P(\Delta u - u \cdot \nabla u)$, we have the estimate

$$\begin{aligned} \|u_t(0)\| &= \|P(\Delta u_0 - u_0 \cdot \nabla u_0)\| \\ &\leq \|\Delta u_0 - u_0 \cdot \nabla u_0\| \leq B(\|u_0\|_{W_2^2(\Omega)}) \end{aligned} \quad [35]$$

provided that u is smooth in $\bar{\Omega} \times [0, T)$. This is a delicate point, having been forewarned of a regularity breakdown at $t=0$. But, we will be able to replicate the estimate [35] for the Galerkin approximations, ultimately validating [34] for the approximations and the solution.

The integration of the next differential inequality, which arises from the second of the identities [31], requires that $\|\nabla u_t(0)\| < \infty$. Similarly to [35], we have

$$\begin{aligned} \|\nabla u_t(0)\| &= \|\nabla P(\Delta u_0 - u_0 \cdot \nabla u_0)\| \\ &\leq B(\|u_0\|_{W_2^3(\Omega)}) \end{aligned} \quad [36]$$

provided that u is smooth in $\bar{\Omega} \times [0, T)$. However, there is a big difference between [35] and [36]. In the next section, we will not be able to obtain an analog of [36] for the Galerkin approximations. Consequently, the solution that is obtained will not be fully regular at time $t=0$. It will satisfy $u \in C(\bar{\Omega} \times [0, T)) \cap C^\infty(\bar{\Omega} \times (0, T))$, but not $u \in C^\infty(\bar{\Omega} \times [0, T))$. It will satisfy

$\|u(t) - u_0\|_{W_2^1(\Omega)} \rightarrow 0$ but not $\|u(t) - u_0\|_{W_2^3(\Omega)} \rightarrow 0$, as $t \rightarrow 0^+$.

One may wonder whether this is a fault or deficiency in the Galerkin method. It is not, remembering what was said at the beginning of this section. For most prescribed values of u_0 , no matter how smooth, there is a breakdown in the regularity of the solution as $t \rightarrow 0^+$. In fact, it was proved in Heywood and Rannacher (1982) that if $\|\nabla u_t(t)\|$ or any one of several other quantities, including $\|u(t)\|_{W_2^3(\Omega)}$, remains bounded as $t \rightarrow 0^+$, then there exists a solution p_0 of the overdetermined Neumann problem

$$\begin{aligned} -\Delta p_0 &= \nabla \cdot (u_0 \cdot \nabla u_0) \quad \text{in } \Omega \\ \nabla p_0|_{\partial\Omega} &= \Delta u_0|_{\partial\Omega} \end{aligned} \quad [37]$$

Generically speaking, this problem is not solvable, and therefore

$$\limsup_{t \rightarrow 0^+} \|\nabla u_t(t)\| = \infty$$

We mention that under our assumption that u_0 is smooth, the correctly posed Neumann problem, with boundary condition $\partial p_0 / \partial n|_{\partial\Omega} = \Delta u_0 \cdot n|_{\partial\Omega}$, is uniquely solvable for a solution $p_0 \in W_2^1(\Omega)/R$, and $\|\nabla p(t) - \nabla p_0\| \rightarrow 0$, as $t \rightarrow 0^+$; see Heywood and Rannacher (1982).

Since solutions are smooth for $0 < t < T$, the pressure in the Navier–Stokes equations satisfies the overdetermined Neumann problem for all $t \in (0, T)$. So it may seem appropriate to require that the prescribed initial value u_0 be a function for which problem [37] is solvable. We do not agree with that. It is too difficult, if not impossible, to find such functions, except by solving the Navier–Stokes equations. For example, one might think that the condition that [37] should be solvable might be satisfied if $u_0 \in D(\Omega)$, since such functions are zero in a neighborhood of the boundary. In fact, K Masuda has shown that if Ω is a three-dimensional sphere, then the overdetermined Neumann problem [37] is never solvable for nonzero $u_0 \in D(\Omega)$. Hence, the gradient of the initial pressure will have a nonzero tangential component, causing an impulsive tangential acceleration along the boundary.

If we are to use the Navier–Stokes equations to make predictions of the future, we must solve the initial boundary value problem for “man-made” initial values, and accept the fact that there is a momentary breakdown in regularity along the boundary, immediately following the initial time. Thereafter, the solution smooths as “nature” takes over. To prove the reliability of our predictions, we need continuous dependence estimates and error estimates for numerical methods that take into

account this initial breakdown in the regularity. The continuous dependence estimate [14] meets this requirement. So also do the error estimates given in a series of four papers by Rannacher and the author, beginning with Heywood and Rannacher (1982). They were based on the “smoothing” regularity estimates for solutions that are being presented here. We go on with these now, as models for similar estimates for the Galerkin approximations.

Estimating the right-hand side of the second of the identities [31] using [20] and Young’s inequality, and then multiplying through by t , we get the linear differential inequality

$$\begin{aligned} \frac{d}{dt} (t \|\nabla u_t\|^2) + t \|P \Delta u_t\|^2 \\ \leq \|\nabla u_t\|^2 + c (\|\nabla u\|^4 + \|\nabla u\|^2 \\ + \|P \Delta u\|^2) (t \|\nabla u_t\|^2) \end{aligned} \quad [38]$$

for $t \|\nabla u_t\|^2$, with coefficients that are integrable in view of the previous estimates [32], [34], and [35]. Therefore, its integration yields an estimate analogous to [32] of the form

$$\begin{aligned} t \|\nabla u_t(t)\|^2 + \int_0^t t \|P \Delta u_t\|^2 d\tau \\ \leq B(M, t, \|u_0\|_{W_2^2(\Omega)}), \quad \text{for } 0 < t < T \end{aligned} \quad [39]$$

provided its “initial value” is finite. It is, due to the time weight, in the sense that

$$\limsup_{t \rightarrow 0^+} (t \|\nabla u_t(t)\|^2) = 0 \quad [40]$$

This is proved by noting that if the lim sup were positive, then the integral on the left-hand side of [34] would be infinite. Finally, a term $t \|u_{tt}\|^2$ can be included under the integral sign on the left-hand side of [39], because $\|u_{tt}\|$ and $\|P \Delta u_t\|$ are of essentially the same order, being the leading terms in the projection $u_{tt} + P(u_t \cdot \nabla u + u \cdot \nabla u_t) = P \Delta u_t$ of the time differentiated Navier–Stokes equation.

We continue inductively. Estimating the right-hand side of the third of the identities [30] using [12], [20], and Young’s inequality, and then multiplying through by t^2 , we get the linear differential inequality

$$\begin{aligned} \frac{d}{dt} (t^2 \|u_{tt}\|^2) + t^2 \|\nabla u_{tt}\|^2 \\ \leq 2t \|u_{tt}\|^2 + t^2 \|\nabla u_t\|^2 + t^2 \|P \Delta u_t\|^2 \\ + c (\|\nabla u\|^4 + \|\nabla u_t\|^4) (t^2 \|u_{tt}\|^2) \end{aligned} \quad [41]$$

with coefficients that are integrable in view of preceding estimates. In particular, the integrability

of the first term on the right-hand side follows from the boundedness of the integral

$$\int_0^t t \|u_{tt}\|^2 d\tau \quad [42]$$

which, we have pointed out, can be included on the left-hand side of [39]. Finally, notice that the boundedness of the integral [42] implies

$$\limsup_{t \rightarrow 0^+} (t^2 \|u_{tt}(t)\|^2) = 0 \quad [43]$$

Therefore, we can integrate [41] to get the estimate

$$\begin{aligned} t^2 \|u_{tt}(t)\|^2 + \int_0^t t^2 \|\nabla u_{tt}\|^2 d\tau \\ \leq B(M, t, \|u_0\|_{W_2^2(\Omega)}), \quad \text{for } 0 \leq t < T \end{aligned} \quad [44]$$

analogous to [34].

At this point, we have introduced every device needed to proceed by induction to an infinite sequence of time-weighted estimates, similar to [39] and [44], but with successively higher orders of time derivatives and weights. The dependence of these estimates on $\|u_0\|_{W_2^2(\Omega)}$ was introduced through [34] and [35]. It can be eliminated by beginning the introduction of powers of t as weight functions one step earlier, with the added advantage that the initial velocity u_0 needs only belong to $J_1(\Omega)$. In the two-dimensional case, the weight functions can be introduced even another step earlier, with the advantage that the initial velocity u_0 needs only belong to $J(\Omega)$. Each of these cases leads to an existence theorem for solutions $u \in C^\infty(\bar{\Omega} \times (0, T))$, with the initial values assumed in the norms of $J_1(\Omega)$ and $J(\Omega)$, respectively.

Existence by Galerkin Approximation

Let $\{a^1, a^2, \dots\}$ and $\{\lambda_1, \lambda_2, \dots\}$ denote the eigenfunctions and eigenvalues of the Stokes equations,

$$\begin{aligned} -\Delta a^k + \nabla p &= \lambda_k a^k, \quad \nabla \cdot a^k = 0 \quad \text{in } \Omega \\ a^k|_{\partial\Omega} &= 0 \end{aligned} \quad [45]$$

chosen to be orthonormal in $L^2(\Omega)$. Clearly, $-P\Delta a_k = \lambda_k a^k$, so they are also the eigenfunctions and eigenvalues of the Stokes operator, $-P\Delta$. Using regularity estimates for the Stokes equations, each eigenfunction is known to be C^∞ smooth in $\bar{\Omega}$.

The n th Galerkin approximation for problem [9] is the solution

$$u^n(x, t) = \sum_{k=1}^n c_{kn}(t) a^k(x)$$

of the system of ordinary differential equations

$$\begin{aligned} (u_t^n, a^l) + (u^n \cdot \nabla u^n, a^l) &= (\Delta u^n, a^l) \\ \text{for } l &= 1, 2, \dots, n \end{aligned} \quad [46]$$

satisfying the initial conditions $(u^n(0), a^l) = (u_0, a^l)$, for $l = 1, 2, \dots, n$. Of course, since $(u_t^n, a^l) = \partial c_{ln} / \partial t$ and $(\Delta u^n, a^l) = (P\Delta u^n, a^l) = -\lambda_l c_{ln}$, the differential equations can be written as

$$\frac{d}{dt} c_{ln} = - \sum_{i,j=1}^n c_{in} c_{jn} (a^i \cdot \nabla a^j, a^l) - \lambda_l c_{ln}$$

and the initial conditions as $c_{ln}(0) = (u^n(0), a^l)$, for $l = 1, 2, \dots, n$.

The system [46] is at least locally solvable, on some interval $[0, T_n)$, with each coefficient satisfying $c_{ln} \in C^\infty[0, T_n)$. Therefore, since the eigenfunctions are also smooth, u^n is C^∞ smooth in $\bar{\Omega} \times [0, T_n)$. It also satisfies all of the identities [30] and [31] on the interval $[0, T_n)$. Indeed, multiplying [46] by c_{ln} and summing over l from 1 to n has the effect of converting a^l into u^n . The resulting identity for u^n leads immediately to the energy identity

$$\frac{1}{2} \frac{d}{dt} \|u^n\|^2 + \|\nabla u^n\|^2 = 0 \quad [47]$$

The remaining identities in the sequence [30] are obtained similarly. For example, the second is obtained by taking the time derivative of [46], multiplying through by dc_{ln}/dt and summing over l .

Prodi's identity is obtained by multiplying [46] by $\lambda_l c_{ln}$ and summing, which has the effect of converting a^l into $-P\Delta u^n$. To obtain the second of the identities [31] for u^n , one differentiates [46], multiplies by $\lambda_l dc_{ln}/dt$ and sums. The remaining identities in the sequence [31] are obtained similarly.

The initial conditions easily imply that $\|u^n(0)\| \leq \|u_0\|$, because $u_0 \in J(\Omega)$ and the eigenfunctions are orthogonal and complete in $J(\Omega)$. Therefore, integration of [47] yields the energy estimate

$$\frac{1}{2} \|u^n(t)\|^2 + \int_0^t \|\nabla u^n\|^2 d\tau \leq \frac{1}{2} \|u_0\|^2 \quad [48]$$

which is uniform in n . Since $\|u^n(t)\|$ remains bounded, the solution $u^n(t)$ can be continued for all time. Thus, $T_n = \infty$, for all n . Hence, our early working assumption about solutions, that they are smooth in $\bar{\Omega} \times [0, \infty)$, is actually valid for the Galerkin approximations. The issue becomes one of obtaining estimates for their derivatives that are uniform in n . All of the estimates we have proved for solutions are proved in exactly the same way for the approximations. The only possible source of nonuniformity would arise from the initial values of $\|\nabla u^n\|$ and $\|u_t^n\|$.

The estimates [24], [26], and [27] are uniform in n , since $u_0 \in J_1(\Omega)$ and hence $\|\nabla u^n(0)\| \leq \|\nabla u_0\|$, due to the orthogonality of the eigenfunctions in the inner-product $(\nabla u, \nabla v)$, and their completeness with respect to functions in $J_1(\Omega)$. We also obtain a uniform bound for $\|u_t^n(0)\|$ of the form [35], by multiplying [46] by $\partial c_{ln}/\partial t$ and summing over l . In the last step, we also need the inequality $\|u^n(0)\|_{W_2^1(\Omega)} \leq \|u_0^n\|_{W_2^1(\Omega)}$, which follows from the orthogonality of the eigenfunctions in the inner product $(P\Delta u, P\Delta v)$, and their completeness with respect to functions in $J_1(\Omega) \cap W_2^2(\Omega)$; see Ladyzhenskaya (1969, p. 46). Any attempt to find a bound for $\|\nabla u_t^n(0)\|$ analogous to [36] is certain to fail, as it would lead to a contradiction with aforementioned results from Heywood and Rannacher (1982).

Passage to the Limit

We now have L^2 -bounds for $u^n, \nabla u^n, u_t^n, \partial^2 u^n / \partial x_i \partial x_j$, and ∇u_t^n over any space-time region $\Omega \times (0, T')$, with $0 < T' < T$. We also have L^2 -bounds for all orders of the time derivatives of these quantities over any subregion $\Omega \times (\varepsilon, T')$, with $0 < \varepsilon < T' < T$. From these L^2 -bounds, we may infer the existence of a subsequence of the Galerkin approximations, again denoted by $\{u^n\}$, which converges, along with those of its derivatives for which we have bounds, to a limit u and its derivatives. The convergence $u^n \rightarrow u$ and $\nabla u^n \rightarrow \nabla u$ is strong in $L^2(\Omega \times (0, T'))$; the convergence of u_t^n is strong in $L^2(\Omega \times (\varepsilon, T'))$ and weak in $L^2(\Omega \times (0, T'))$; the convergence of $P\Delta u^n$ is weak in $L^2(\Omega \times (0, T'))$; all time derivatives of $u^n, \nabla u^n$ converge strongly in $L^2(\Omega \times (\varepsilon, T'))$.

Because of estimates for the time derivatives, trace arguments give the strong convergence $u^n \rightarrow u$, $\nabla u^n \rightarrow \nabla u$, $u_t^n \rightarrow u_t$, and the weak convergence $P\Delta u^n \rightarrow P\Delta u$, in $L^2(\Omega)$, for every $t > 0$.

For any fixed time, $u \in W_2^1(\Omega)$, and therefore u is continuous in $\bar{\Omega}$ by a well known Sobolev inequality. Since $u \in J_1(\Omega)$, it must equal zero along the boundary. The estimates for the time derivatives of $u^n, \nabla u^n, \partial^2 u^n / \partial x_i \partial x_j$ imply that u and its time derivatives are time continuous in $W_2^1(\Omega)$. Therefore, u, u_t, u_{tt}, \dots are classically continuous in $\bar{\Omega} \times (0, T)$.

Introduction of the Pressure

Because of the strong convergence $u^n \rightarrow u$, $\nabla u^n \rightarrow \nabla u$, $u_t^n \rightarrow u_t$ and the weak convergence $P\Delta u^n \rightarrow P\Delta u$, in $L^2(\Omega)$, for any $t > 0$, it is an easy matter to let $n \rightarrow \infty$ in [46], obtaining, for all $t > 0$,

$$\begin{aligned} (u_t, a^l) + (u \cdot \nabla u, a^l) &= (\Delta u, a^l) \\ \text{for } l &= 1, 2, \dots \end{aligned} \quad [49]$$

Since the eigenfunctions are complete in $J(\Omega)$, and $D(\Omega) \subset J(\Omega)$, this implies

$$(u_t + u \cdot \nabla u - \Delta u, \phi) = 0, \quad \text{for all } \phi \in D(\Omega) \quad [50]$$

Therefore, there exists a vector field $\nabla p \in G(\Omega)$ such that

$$u_t + u \cdot \nabla u - \Delta u = -\nabla p \quad [51]$$

Indeed, the usual test to determine whether a smooth vector field w is conservative in some domain Ω , and therefore representable as a gradient, is to check whether the curve integrals

$$\oint_C w \cdot \tau \, ds \quad [52]$$

vanish for every smooth closed curve $C \subset \Omega$. Here, τ is the unit tangent to the curve and ds is its arc length. With a little reflection, one will realize that these curve integrals can be approximated by volume integrals of the form (w, ϕ) with $\phi \in D(\Omega)$. For this, one should choose ϕ to have its support in a small tubular neighborhood of the curve, and its streamlines parallel to the curve, with unit net flux through any section of the tube. If w is not smooth, but only known to belong to $L^2(\Omega)$, one can approximate it with its smooth mollifications. This argument can be made rigorous. We previously showed that $J(\Omega)$ and $G(\Omega)$ are orthogonal subspaces of $L^2(\Omega)$. Now we have argued that $L^2(\Omega) = J(\Omega) \oplus G(\Omega)$.

Classical C^∞ Regularity

At any fixed time, we may regard u as a solution of the steady Stokes problem [18] with $f = -u_t - u \cdot \nabla u$. Included in Cattabriga (1961) and Solonnikov (1964, 1966) are regularity estimates for all orders of derivatives of the form

$$\|u\|_{W_2^{k+2}(\Omega)} \leq c \|f\|_{W_2^k(\Omega)}$$

From our estimates above, we easily conclude that $f \equiv -u_t - u \cdot \nabla u \in W_2^1(\Omega)$. Hence, $u \in W_2^3(\Omega)$. In fact, in view of the regularity we have proven with respect to time, $f \in C^\infty(0, T; W_2^1(\Omega))$ and $u \in C^\infty(0, T; W_2^3(\Omega))$. Thus begins a bootstrapping argument. In the next step, we observe that $f \in C^\infty(0, T; W_2^2(\Omega))$ and conclude that $u \in C^\infty(0, T; W_2^4(\Omega))$. By induction, one obtains $u \in C^\infty(0, T; W_2^k(\Omega))$ for every positive integer k . Then well-known Sobolev inequalities imply that $u \in C^\infty(\bar{\Omega} \times (0, T))$.

Assumption of the Initial Values

We begin by showing that $u(t) \rightarrow u_0$, weakly in $L^2(\Omega)$, as $t \rightarrow 0^+$. Of course, $\|u(t)\|$ remains bounded as $t \rightarrow 0^+$, in virtue of [48], and the eigenfunctions $\{a^l\}$ are complete in $J(\Omega)$. Writing

$$\begin{aligned} (u(t) - u_0, a^l) &= (u(t) - u^n(t), a^l) + (u^n(t) \\ &\quad - u^n(0), a^l) + (u^n(0) - u_0, a^l) \end{aligned}$$

note that the first and third terms on the right-hand side can be made small by choosing n large. The second can be written as

$$(u^n(t) - u^n(0), a^l) = \int_0^t (u_t^n, a^l) d\tau$$

which will be small if t is small, in view of [34]. Thus, $(u(t) - u_0, a^l) \rightarrow 0$, as $t \rightarrow 0^+$, which implies the desired weak convergence.

The strong convergence $u(t) \rightarrow u_0$ in $L^2(\Omega)$ follows from the weak convergence if $\limsup_{t \rightarrow 0^+} \|u(t)\| \leq \|u_0\|$. The energy estimate [48] for the approximations implies this also.

To conclude that $u(t) \rightarrow u_0$ strongly in $J_1(\Omega)$, it only remains to be shown that $\limsup_{t \rightarrow 0^+} \|\nabla u(t)\| \leq \|\nabla u_0\|$. This readily follows from [29], provided the bounding function $M(t)$ satisfies $M(t) \rightarrow \|\nabla u_0\|$, as $t \rightarrow 0^+$. The bounding functions provided by our basic estimates [24], [26], and [27] all have this property.

We may conclude that $u(t) \rightarrow u_0$ weakly in $W_2^2(\Omega)$, provided $\|u(t)\|_{W_2^2(\Omega)}$ remains bounded as $t \rightarrow 0^+$. To see this, remember that $\|P\Delta u\|$ and $\|u_t\|$ are of essentially the same order. Thus the term $\|u_t(t)\|^2$ on the left-hand side of [34] can be accompanied by a term $\|u(t)\|_{W_2^2(\Omega)}^2$.

Finally, to prove that $u(t) \rightarrow u_0$ strongly in $W_2^2(\Omega)$, we need only show that $\limsup_{t \rightarrow 0^+} \|P\Delta u(t)\| \leq \|P\Delta u_0\|$, since $\|P\Delta \cdot\|$ and $\|\cdot\|_{W_2^2(\Omega)}$ are equivalent norms on $J_1(\Omega) \cap W_2^2(\Omega)$. To this end, multiply [46] by $\lambda_l dc_{ln}/dt$ and sum to get

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|P\Delta u^n\|^2 + \|\nabla u_t^n\|^2 &= (u^n \cdot \nabla u^n, P\Delta u_t^n) \\ &= \frac{d}{dt} (u^n \cdot \nabla u^n, P\Delta u^n) \\ &\quad - (u_t^n \cdot \nabla u^n + u^n \cdot \nabla u_t^n, P\Delta u^n) \end{aligned}$$

Integrating this gives

$$\begin{aligned} \|P\Delta u^n(t)\|^2 - \|P\Delta u^n(0)\|^2 &= \int_0^t \frac{d}{ds} \|P\Delta u^n\|^2 ds \\ &= 2(u^n \cdot \nabla u^n, P\Delta u^n)|_t - 2(u^n \cdot \nabla u^n, P\Delta u^n)|_0 \\ &\quad - 2 \int_0^t \|\nabla u_t^n\|^2 ds - 2 \int_0^t (u_t^n \cdot \nabla u^n \\ &\quad + u^n \cdot \nabla u_t^n, P\Delta u^n) ds \end{aligned} \quad [53]$$

For the terms under the last integral we have

$$\begin{aligned} |(u_t^n \cdot \nabla u^n, P\Delta u^n) + (u^n \cdot \nabla u_t^n, P\Delta u^n)| \\ \leq \|\nabla u_t^n\|^2 + c \|\nabla u^n\|^{1/2} \|P\Delta u^n\|^{3/2} \end{aligned}$$

Therefore, [53] implies

$$\begin{aligned} \|P\Delta u^n(t)\|^2 \leq \|P\Delta u^n(0)\|^2 + 2(u^n \cdot \nabla u^n, P\Delta u^n)|_t \\ - 2(u^n \cdot \nabla u^n, P\Delta u^n)|_0 + Kt \end{aligned}$$

uniformly in n , as $t \rightarrow 0^+$, where K is a constant depending on the estimates [32] and [34]. Letting $n \rightarrow \infty$, gives

$$\begin{aligned} \|P\Delta u(t)\|^2 \leq \|P\Delta u(0)\|^2 + 2(u \cdot \nabla u, P\Delta u)|_t \\ - 2(u \cdot \nabla u, P\Delta u)|_0 + Kt \end{aligned}$$

Since $u \cdot \nabla u \rightarrow u_0 \cdot \nabla u_0$ strongly in $L^2(\Omega)$, and $P\Delta u \rightarrow P\Delta u_0$ weakly in $L^2(\Omega)$, we get the desired result. The continuous assumption of the initial values in $W_2^2(\Omega)$ also implies their continuous assumption in the classical sense, and hence that $u \in C(\bar{\Omega} \times [0, T])$.

Conclusion

Years ago, mathematical questions concerning the Navier–Stokes equations were usually considered in the context of generalized or weak solutions, which was a technical barrier to many in the scientific community. Nowadays, realizing that solutions are at least locally classical, fundamental questions such as that of global regularity can be studied within the classical context. If the estimate [29] is proved for classical solutions, with $T = \infty$, and without a restriction on the size of the data, this particular matter will be settled.

Acknowledgment

This work has been supported by the Natural Sciences and Engineering Research Council of Canada.

See also: Compressible Flows: Mathematical Theory; Elliptic Differential Equations: Linear Theory; Incompressible Euler Equations: Mathematical Theory; Interfaces and Multicomponent Fluids; Leray–Schauder Theory and Mapping Degree; Non-Newtonian Fluids; Partial Differential Equations: Some Examples; Stochastic Hydrodynamics; Turbulence Theories; Wavelets: Application to Turbulence.

Further Reading

- Beirão da Veiga H (1997) A new approach to the L^2 -regularity theorem for linear stationary nonhomogeneous Stokes systems. *Portugaliae Mathematica* 54(Fasc. 3): 271–286.
- Cattabriga L (1961) Su un problema al contorno relativo al sistema di equazioni di Stokes. *Rendi conti del Seminario Matematico della Università di Padova* 31: 308–340.

- Heywood JG (1976) On uniqueness questions in the theory of viscous flow. *Acta Mathematica* 136: 61–102.
- Heywood JG (1980) The Navier–Stokes equations: on the existence, regularity and decay of solutions. *Indiana University Mathematics Journal* 29: 639–681.
- Heywood JG (1990) Open problems in the theory of the Navier–Stokes equations for viscous incompressible flow. In: Heywood JG, Masuda K, Rautmann R, and Solonnikov VA (eds.) *The Navier–Stokes Equations: Theory and Numerical Methods*, Lecture Notes in Mathematics, vol. 1431, pp. 1–22. Berlin–Heidelberg: Springer Verlag.
- Heywood JG (1994) Remarks concerning the possible global regularity of solutions of the three-dimensional incompressible Navier–Stokes equations. In: Galdi GP, Malek J, and Necas J (eds.) *Progress in Theoretical and Computational Fluid Dynamics*, Pitman Research Notes in Mathematics Series, vol. 308, pp. 1–32. Essex: Longman Scientific and Technical.
- Heywood JG (2001) On a conjecture concerning the Stokes problem in nonsmooth domains. In: Neustupa J and Penel P (eds.) *Mathematical Fluid Mechanics: Recent Results and Open Problems*, Advances in Mathematical Fluid Mechanics vol. 2, pp. 195–206. Basel: Birkhauser Verlag.
- Heywood JG (2003) A curious phenomenon in a model problem, suggestive of the hydrodynamic inertial range and the smallest scale of motion. *Journal of Mathematical Fluid Mechanics* 5: 403–423.
- Heywood JG and Rannacher R (1982) Finite element approximation of the nonstationary Navier–Stokes problem, Part I: Regularity of solutions and second-order error estimates for the spatial discretization. *SIAM Journal on Numerical Analysis* 19: 275–311.
- Heywood JG and Xie W (1997) Smooth solutions of the vector Burgers equation in nonsmooth domains. *Differential and Integral Equations* 10: 961–974.
- Ladyzhenskaya OA (1969) *The Mathematical Theory of Viscous Incompressible Flow*, 2nd edn. New York: Gordon and Breach.
- Prodi G (1962) Teoremi di tipo lacale per il sistema de Navier–Stokes e stabilità delle soluzione stazionarie. *Rendi conti del Seminario Matematico della Università di Padova* 32: 374–397.
- Solonnikov VA (1964) On general boundary-value problems for elliptic systems in the sense of Douglis–Nirenberg. I. *Izvestiya Akademii Nauk SSSR, Soriya Matematicheskaya* 28: 665–706.
- Solonnikov VA (1966) On general boundary-value problems for elliptic systems in the sense of Douglis–Nirenberg. II. *Trudy Matematicheskogo Instituta Imeni V.A. Steklova* 92: 233–297.
- Solonnikov VA and Ščadilov VE (1973) On a boundary value problem for a stationary system of Navier–Stokes equations. *Proceedings. Steklov Institute of Mathematics* 125: 186–199.
- Xie W (1991) A sharp pointwise bound for functions with L^2 Laplacians and zero boundary values on arbitrary three-dimensional domains. *Indiana University Mathematics Journal* 40: 1185–1192.
- Xie W (1992) On a three-norm inequality for the Stokes operator in nonsmooth domains. In: Heywood JG, Masuda K, Rautmann R, and Solonnikov VA (eds.) *The Navier–Stokes Equations II: Theory and Numerical Methods*, Lecture Notes in Mathematics, vol. 1530, pp. 310–315. Berlin–Heidelberg: Springer Verlag.
- Xie W (1997) Sharp Sobolev interpolation inequalities for the Stokes operator. *Differential and Integral Equations* 10: 393–399.

von Neumann Algebras: Introduction, Modular Theory, and Classification Theory

V S Sunder, The Institute of Mathematical Sciences, Chennai, India

© 2006 Elsevier Ltd. All rights reserved.

Introduction

von Neumann algebras, as they are called now, first made their appearance under the name “rings of operators” in a series of seminal papers – see Murray and von Neumann (1936, 1937, 1943) and von Neumann (1936) – by F J Murray and J von Neumann starting in 1936. Murray and von Neumann (1936) specifically cite “attempts to generalize the theory of unitary group representations” and “demands by various aspects of the quantum-mechanical formalism” among the reasons for the elucidation of this subject.

In fact, the simplest definition of a von Neumann algebra is via unitary group representations: a collection M of continuous linear operators on a Hilbert space \mathcal{H} (in order to avoid some potential technical problems, we shall restrict ourselves to

separable Hilbert spaces throughout this article) is a von Neumann algebra precisely when there is a representation ρ of a group G as unitary operators on \mathcal{H} such that

$$M = \{x \in \mathcal{L}(\mathcal{H}) : x\rho(t) = \rho(t)x \ \forall t \in G\}$$

As above, we shall write $\mathcal{L}(\mathcal{H})$ for the collection of all continuous linear operators on the Hilbert space \mathcal{H} ; recall that a linear mapping $x : \mathcal{H} \rightarrow \mathcal{H}$ is continuous precisely when there exists a positive constant K such that $\|x\xi\| \leq K\|\xi\| \ \forall \xi \in \mathcal{H}$. If the norm $\|x\|$ of the operator x is defined as the smallest constant K with the above property, then the set $\mathcal{L}(\mathcal{H})$ acquires the structure of a Banach space. In fact $\mathcal{L}(\mathcal{H})$ is a Banach $*$ -algebra with respect to the composition product, and involution $x \mapsto x^*$ given by

$$\langle x\xi, \eta \rangle = \langle \xi, x^*\eta \rangle \ \forall \xi, \eta \in \mathcal{H}$$

The first major result in the subject is the remarkable “double commutant theorem,” which establishes the equivalence of a purely algebraic requirement to purely topological ones. We need

two bits of terminology to be able to state the theorem.

First, define the commutant S' of a subset $S \subset \mathcal{L}(\mathcal{H})$ by

$$S' = \{x' \in \mathcal{L}(\mathcal{H}) : x'x = xx' \forall x \in S\}$$

Second, the strong (resp., weak) operator topology is the topology on $\mathcal{L}(\mathcal{H})$ of “pointwise strong (resp., weak) convergence”: that is, $x_n \rightarrow x$ precisely when $\|x_n\xi - x\xi\| \rightarrow 0 \forall \xi \in \mathcal{H}$ (resp., $\langle x_n\xi - x\xi, \eta \rangle \rightarrow 0 \forall \xi, \eta \in \mathcal{H}$).

Theorem 1 *The following conditions on a unital $*$ -subalgebra M of $\mathcal{L}(\mathcal{H})$ are equivalent:*

- (i) $M = M'' (= (M')')$.
- (ii) M is closed in the strong operator topology.
- (iii) M is closed in the weak operator topology.

The conventional definition of a von Neumann algebra is that it is a unital $*$ -subalgebra of $\mathcal{L}(\mathcal{H})$ which satisfies the equivalent conditions above. The equivalence with our earlier “simplest definition” is a consequence of the double commutant theorem and the fact that any element of a von Neumann algebra is a linear combination of four unitary elements of the algebra: simply take G to be the group of unitary operators in M' .

Another consequence of the double commutant theorem is that von Neumann algebras are closed under any “canonical construction.” For instance, the uniqueness of the spectral measure $E \mapsto P_x(E)$ associated to a normal operator x shows that if u is unitary, then $P_{uxu^*}(E) = uP_x(E)u^*$ for all Borel sets E . In particular, if $x \in M$ and $u' \in \mathcal{U}(M')$, then $u'P_x(E)u'^* = P_{u'xu'^*}(E) = P_x(E)$, and hence, we may conclude that $P_x(E) \in \mathcal{U}(M')' = (M')' = M$ (we will write $\mathcal{U}(N)$ (resp., $\mathcal{P}(N)$) to denote the collection of unitary (resp., projection) operators in any von Neumann algebra N); that is, if a von Neumann algebra contains a normal operator, it also contains all the associated spectral projections. This fact, together with the spectral theorem, has the consequence that any von Neumann algebra M is the closed linear span of $\mathcal{P}(M)$.

The analogy with unitary group representations is fruitful. Suppose then that $M = \rho(G)'$, for a unitary representation of G . Then the last sentence of the previous paragraph implies that $\rho(G)' = \mathbb{C}$ precisely when there exist no nontrivial ρ -stable subspaces (here and in the sequel, we identify \mathbb{C} with its image under the unique unital homomorphism of \mathbb{C} into $\mathcal{L}(\mathcal{H})$; and we reserve the symbol $Z(M)$ to denote the center of M), that is, when ρ is irreducible. In general, the ρ -stable subspaces are precisely the ranges of projection operators in M . The notion of unitary

equivalence of subrepresentations of ρ is seen to translate to the equivalence defined on the set $\mathcal{P}(M)$ of projections in M , whereby $p \sim q$ if and only if there exists an operator $u \in M$ such that $u^*u = p$ and $uu^* = q$. (Such a u is called a partial isometry, with “initial space” = range p , and “final space” = range q .) This is the definition of what is known as the “Murray–von Neumann equivalence rel M ” and is denoted by \sim_M . The following accompanying definition is natural: if $p, q \in \mathcal{P}(M)$, say $p \preceq_M q$ if there exists $p_0 \in \mathcal{P}(M)$ such that $p \sim_M p_0 \leq q$ – where of course $e \leq f \Leftrightarrow \text{range}(e) \subset \text{range}(f)$.

The Murray–von Neumann Classification of Factors

We start with a fact (whose proof is quite easy) and a consequent fundamental definition.

Proposition 2 *The following conditions on a von Neumann algebra M are equivalent:*

- (i) for any $p, q \in \mathcal{P}(M)$, it is true that either $p \preceq_M q$ or $q \preceq_M p$.
- (ii) $Z(M) = M \cap M' = \mathbb{C}$.

The von Neumann algebra M is called a “factor” if it satisfies the equivalent conditions above.

The alert reader would have noticed that if G is a finite group, then $\rho(G)'$ is a factor precisely when the representation ρ is “isotypical.” Thus, the “representation-theoretic fact,” that any unitary representation is expressible as a direct sum of isotypical subrepresentations, translates into the “von Neumann algebraic fact” that any $*$ -subalgebra of $\mathcal{L}(\mathcal{H})$ is isomorphic, when \mathcal{H} is finite dimensional, to a direct sum of factors. In complete generality, von Neumann (1949) showed that any von Neumann algebra is expressible as a “direct integral of factors.” We shall interpret this fact from “reduction theory” as the statement that all the magic/mystery of von Neumann algebras is contained in factors and hence restrict ourselves, for a while, to the consideration of factors.

Murray and von Neumann initiated the study of a general factor M via a qualitative as well as a quantitative analysis of the relation \preceq_M on $\mathcal{P}(M)$. First, call a $p \in \mathcal{P}(M)$ infinite if there exists a $p_0 \leq p$ such that $p \sim_M p_0$ and $p_0 \neq p$; otherwise, say p is finite. They obtained an analog, called the “dimension function,” of the Haar measure, as follows.

Theorem 3

- (i) *With M as above, there exists a function $D_M : \mathcal{P}(M) \rightarrow [0, \infty]$ which satisfies the following*

properties, and is determined up to a multiplicative constant, by them:

- $p \preceq_M q \Leftrightarrow D_M(p) \leq D_M(q)$
- p is finite if and only if $D_M(p) < \infty$
- If $\{p_n : n = 1, 2, \dots\}$ is any sequence of pairwise orthogonal projections in $\mathcal{P}(M)$ and $p = \sum_n p_n$, then $D_M(p) = \sum_n D_M(p_n)$

(ii) M falls into exactly one of five possible cases, depending on which of the following sets is the range of some scaling of D_M :

- $(I_n) \{0, 1, 2, \dots, n\}$
- $(I_\infty) \{0, 1, 2, \dots, \infty\}$
- $(II_1) [0, 1]$
- $(II_\infty) [0, \infty]$
- $(III) \{0, \infty\}$

In words, we may say that a factor M is of:

1. type I (i.e., of type I_n for some $1 \leq n \leq \infty$) precisely when M contains a minimal projection,
2. type II (i.e., of type II_1 or II_∞) precisely when M contains nonzero finite projections but no minimal projections, and
3. type III precisely when M contains no nonzero finite projections.

Examples $L^\infty(\Omega, \mu)$ may be regarded as a von Neumann algebra acting on $L^2(\Omega, \mu)$ as multiplication operators; thus, if we set $m_f(\xi) = f\xi$, then $m : f \mapsto m_f$ defines an isomorphism of $L^\infty(\Omega, \mu)$ onto a commutative von Neumann subalgebra of $\mathcal{L}(L^2(\Omega, \mu))$. In fact, “up to multiplicity,” this is how any commutative von Neumann algebra looks.

It is a simple exercise to prove that $M \subset \mathcal{L}(\mathcal{H})$ is a factor of type I_n , $1 \leq n \leq \infty$, if and only if there exist Hilbert spaces \mathcal{H}_n and \mathcal{K} and identifications $\mathcal{H} = \mathcal{H}_n \otimes \mathcal{K}$, $M = \{x \otimes \text{id}_{\mathcal{K}} : x \in \mathcal{L}(\mathcal{H}_n)\}$ where $\dim \mathcal{H}_n = n$; and so $M \cong \mathcal{L}(\mathcal{H}_n)$.

To discuss examples of the other types, it will be convenient to use “crossed products” of von Neumann algebras by ergodically acting groups of automorphisms. We shall now digress with a discussion of this generalization of the notion of a semidirect product of groups.

If $\alpha : G \rightarrow \text{Aut}(M)$ is an action of a countable group G on M , where $M \subset \mathcal{L}(\mathcal{H})$ is a von Neumann algebra, and $\tilde{\mathcal{H}} = \ell^2(G, \mathcal{H})$, there are representations $\pi : M \rightarrow \mathcal{L}(\tilde{\mathcal{H}})$ and $\lambda : G \rightarrow \mathcal{U}(\mathcal{L}(\tilde{\mathcal{H}}))$ defined by

$$(\pi(x)\xi)(s) = \alpha_s(x)\xi(s), \quad (\lambda(t)\xi)(s) = \xi(t^{-1}s)$$

These representations satisfy the commutation relation $\lambda(t)\pi(x)\lambda(t^{-1}) = \pi(\alpha_t(x))$, and the crossed product $M \rtimes_\alpha G$ is the von Neumann subalgebra of $\mathcal{L}(\tilde{\mathcal{H}})$ defined by $\tilde{M} = (\pi(M) \cup \lambda(G))''$.

Let us restrict ourselves to the case of $M = L^\infty(\Omega, \mu)$ acting on $L^2(\Omega, \mu)$. In this case, it is true that any automorphism of M is of the form $f \mapsto f \circ T^{-1}$, where T is a “nonsingular transformation of the measure space (Ω, μ) ” (= a bijection which preserves the class of sets of μ -measure 0). So, an action of G on M is of the form $\alpha_t(f) = f \circ T_t^{-1}$, for some homomorphism $t \mapsto T_t$ from G to the group of nonsingular transformations of (Ω, μ) . We have the following elegantly complete result from Murray and von Neumann (1936).

Theorem 4 Let M, G, α be as in the last section, and let $\tilde{M} = M \rtimes_\alpha G$. Assume the G -action is “free,” meaning that if $t \neq 1 \in G$, then $\mu(\{\omega \in \Omega : T_t(\omega) = \omega\}) = 0$. Then:

- (i) \tilde{M} is a factor if and only if G acts ergodically on (Ω, μ) – meaning that the only G -fixed functions in M are the constants.
- (ii) Assume that G acts ergodically. Then the type of the factor \tilde{M} is determined as follows:
 - \tilde{M} is of type I or II if and only if there exists a G -invariant measure ν which is mutually absolutely continuous with respect to μ , meaning $\nu(E) = 0 \Leftrightarrow \mu(E) = 0$; (the ergodicity assumption implies that such a ν is necessarily unique up to scaling by a positive constant;)
 - \tilde{M} is of type I_n precisely when the ν as above is totally atomic, and Ω is the disjoint union of n atoms for ν ;
 - \tilde{M} is of type II precisely when the ν as above is nonatomic;
 - \tilde{M} is of finite type – meaning that 1 is a finite projection in \tilde{M} – precisely when the ν as above is a finite measure;
 - \tilde{M} is of type III if and only if there exists no ν as above.

Thus, we get all the types of factors by this construction; for instance, we may take:

- $(I_n) G = \mathbb{Z}_n$ acting on $\Omega = \mathbb{Z}_n$ by translation, and $\mu = \nu =$ counting measure
- $(I_\infty) G = \mathbb{Z}$ acting on $\Omega = \mathbb{Z}$ by translation, and $\mu = \nu =$ counting measure
- $(II_1) G = \mathbb{Z}$ acting on $\Omega = \mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ by powers of an aperiodic rotation, and $\mu = \nu =$ arclength measure
- $(II_\infty) G = \mathbb{Q}$ acting on $\Omega = \mathbb{R}$ by translations, and $\mu = \nu =$ Lebesgue measure
- $(III) G = ax + b$ group acting in the obvious manner on $\Omega = \mathbb{R}$, $\mu = \nu =$ Lebesgue measure.

Such crossed products of a commutative von Neumann algebra by an ergodically acting countable

group were intensively studied by Krieger (1970, 1976). We shall simply refer to such factors as “Krieger factors.” The term “Krieger factor” is actually used for factors obtained from a slightly more general construction, with ergodic group actions replaced by more general ergodic equivalence relations. Since there is no difference in the two notions at least in good (amenable) cases, we will say no more about this.

Abstract von Neumann Algebras

So far, we have described matters as they were in von Neumann’s time. To come to the modern era, it is desirable to “free a von Neumann algebra from the ambient Hilbert space” and to regard it as an abstract object in its own right which can act on different Hilbert spaces – for example, $L^\infty(\Omega, \mu)$ is an object worthy of study in its own right, without reference to $L^2(\Omega, \mu)$.

The abstract viewpoint is furnished by a theorem of Sakai (1983); let us define an abstract von Neumann algebra to be an abstract C^* -algebra (this is a Banach algebra with an involution related to the norm by the so-called C^* -identity $\|x\|^2 = \|x^*x\|$) M which admits a pre-dual M_* – i.e., M is isometrically isomorphic to the Banach dual space $(M_*)^*$. It turns out that a predual of such an abstract von Neumann algebra is unique up to isometric isomorphism. Consequently, an abstract von Neumann algebra comes equipped with a canonical “weak*-topology,” usually called the “ σ -weak topology” on M . The natural morphisms in the category of abstract von Neumann algebras are $*$ -homomorphisms which are continuous with respect to σ -weak topologies on domain and range. It is customary to call a linear map between abstract von Neumann algebras “normal” if it is continuous with respect to σ -weak topologies on domain and range.

The equivalence of the “abstract” definition of this section, with the “concrete” one given earlier (which depends on an ambient Hilbert space), relies on the following four facts:

1. $\mathcal{L}(\mathcal{H})$ is an abstract von Neumann algebra, with the predual $\mathcal{L}(\mathcal{H})_*$ being the so-called “trace class” of operators, equipped with the “trace norm.”
2. A self-adjoint subalgebra of $\mathcal{L}(\mathcal{H})$ is closed in the strong operator topology, and is hence a “concrete von Neumann algebra” precisely when it is closed in the σ -weak topology on $\mathcal{L}(\mathcal{H})$.
3. If M is an abstract von Neumann algebra, and N is a $*$ -subalgebra of M which is closed in the σ -weak topology of M , then N is also an abstract

von Neumann algebra, with one candidate for N_* being M_*/N_\perp (where $N_\perp = \{\rho \in M_* : n(\rho) = 0 \ \forall n \in N\}$).

4. Any abstract von Neumann algebra (with separable predual) is isomorphic (in the category of abstract von Neumann algebras) to a (concrete) von Neumann subalgebra of $\mathcal{L}(\mathcal{H})$ (for a separable \mathcal{H}).

With the abstract viewpoint available, we shall look for modules over a von Neumann algebra M , meaning pairs (\mathcal{H}, π) where $\pi: M \rightarrow \mathcal{L}(\mathcal{H})$ is a normal $*$ -homomorphism.

A brief digression into the proof of fact (4) above – which asserts the existence of faithful M -modules – will be instructive and useful. Suppose M is an abstract von Neumann algebra. A linear functional ϕ on M is called a normal state if:

- (positivity) $\phi(x^*x) \geq 0 \ \forall x \in M$;
- (normality) $\phi: M \rightarrow \mathbb{C}$ is normal; and
- (normalization) $\phi(1) = 1$.

(Normal states on $L^\infty(\Omega, \mu)$ correspond to non-negative probability measures on Ω which are absolutely continuous with respect to μ .) It is true that there exist plenty of normal states on M . In fact, they linearly span M_* . This implies that if M_* is separable, then there exist normal states on M which are even “faithful” – meaning $\phi(x^*x) = 0 \Leftrightarrow x = 0$.

Fix a faithful normal state ϕ on M . (Consistent with our convention about separable \mathcal{H} ’s, we shall only consider M ’s with separable preduals.) The well-known “Gelfand–Naimark–Segal” construction then yields a faithful M -module which is usually denoted by $L^2(M, \phi)$ – motivated by the fact that if $M = L^\infty(\Omega, \mu)$, and $\phi(f) = \int f d\nu$, with ν a probability measure mutually absolutely continuous with respect to μ , then $L^2(M, \phi) = L^2(\Omega, \nu)$ with $L^\infty(\Omega, \mu)$ acting as multiplication operators. The construction mimics this case: the assumptions on ϕ ensure that the equation

$$\langle x, y \rangle = \phi(y^*x)$$

defines a positive-definite inner product on M ; let $L^2(M, \phi)$ be the Hilbert space completion of M . It turns out that the operator of left-multiplication by an element of M extends as a bounded operator to $L^2(M, \phi)$, and it then follows easily that $L^2(M, \phi)$ is indeed a faithful M -module, thereby establishing fact (4) above.

Since we wish to distinguish between elements of the dense subspace M of $L^2(M, \phi)$ and the operators of left-multiplication by members of M , let us write \hat{x} for an element of M when thought of as an

element of $L^2(M, \phi)$, and x for the operator of left-multiplication by x ; thus, for instance, $\hat{x} = x\hat{1}$, and $x\hat{y} = \widehat{xy}$, $\langle x\hat{1}, \hat{1} \rangle = \phi(x)$, etc.

Modular Theory

While type III factors were more or less an enigma at the time of von Neumann, all that changed with the advent of Connes. The first major result of this “type III era” is the celebrated “Tomita–Takesaki theorem” (cf. Takesaki (1970)), which views the adjoint mapping on M as an appropriate operator on $L^2(M, \phi)$, and analyzes its polar decomposition. Specifically, we have:

Theorem 5 *If ϕ is any faithful normal state on M , consider the densely defined conjugate-linear operator given, with domain $\{\hat{x} : x \in M\}$, by $S_\phi^{(0)}(\hat{x}) = \hat{x}^*$. Then,*

- (i) *there is a unique conjugate-linear operator S_ϕ (the “closure of $S_\phi^{(0)}$ ”) whose graph is the closure of the graph of $S_\phi^{(0)}$; if we write $S_\phi = J_\phi \Delta_\phi^{1/2}$ for the polar decomposition of the conjugate-linear closed operator S_ϕ , then*
- (ii) *J_ϕ is an antiunitary involution on $L^2(M, \phi)$ (i.e., it is a conjugate-linear norm-preserving bijection of $L^2(M, \phi)$ onto itself which is its own inverse);*
- (iii) *Δ_ϕ is an injective positive self-adjoint operator on $L^2(M, \phi)$ such that $J_\phi f(\Delta_\phi) J_\phi = \bar{f}(\Delta_\phi^{-1})$ for all Borel functions $f : \mathbb{R} \rightarrow \mathbb{R}$, and most crucially*
- (iv)

$$J_\phi M J_\phi = M' \quad \text{and} \quad \Delta_\phi^{it} M \Delta_\phi^{-it} = M \quad \forall t \in \mathbb{R}$$

(Here and elsewhere, we shall identify $x \in M$ with the operator of “left-multiplication by x ” on $L^2(M, \phi)$.)

Thus, each faithful normal state ϕ on M yields a one-parameter group $\{\sigma_t^\phi : t \in \mathbb{R}\}$ of automorphisms of M – referred to as the group of “modular automorphisms” – given by

$$\sigma_t^\phi(x) = \Delta_\phi^{it} x \Delta_\phi^{-it}$$

The extent of dependence of the modular group on the state is captured precisely by Connes’ Radon–Nikodym theorem (Connes 1973), which shows that the modular groups associated to two different faithful normal states are related by a “unitary cocycle in M .” This has the consequence that if $\epsilon : \text{Aut}(M) \rightarrow \text{Out}(M) = \text{Aut}(M)/\text{Int}(M)$ is the quotient mapping – where $\text{Int}(M)$ denotes the normal subgroup of inner automorphisms given by unitary

elements of M – then the one-parameter subgroup $\{\epsilon(\sigma_t^\phi) : t \in \mathbb{R}\}$ of $\text{Out}(M)$ is independent of ϕ .

Connes’ Classification and Injective Factors

Given a factor M , Connes defined

$$S(M) =$$

$$\bigcap \{\text{spec}(\Delta_\phi) : \phi \text{ a faithful normal state on } M\}$$

which is obviously an isomorphism invariant. He then classified (Connes 1973) type III factors into a continuum of factors:

Theorem 6 *Let M be a factor. Then,*

- (i) $0 \in S(M) \Leftrightarrow M$ is of type III; and
- (ii) *if M is a type III factor, there are three mutually exclusive and exhaustive possibilities:*

- $(III_0)S(M) = \{0, 1\}$
- $(III_\lambda)S(M) = \{0\} \cup \lambda\mathbb{Z}$, for some $0 < \lambda < 1$
- $(III_1)S(M) = [0, \infty)$

Example 7 Consider the compact group $\Omega = \prod_{n=1}^\infty G_n$ where G_n is a finite cyclic group of order ν_n for each n . Let $\mu = \prod_{n=1}^\infty \mu_n$, where μ_n is a probability measure defined on the subsets of G_n which assigns positive mass to each singleton. Let $G = \bigoplus_{n=1}^\infty G_n$ be the dense subgroup of Ω consisting of finitely nonzero sequences. It is not hard to see that each translation $T_g, g \in G$ (given by $T_g(\omega) = g + \omega$) is a nonsingular transformation of the measure space (Ω, μ) . The density of G in Ω shows that this action of G on $L^\infty(\Omega, \mu)$ is fixed-point-free and ergodic, with the result that the crossed product $L^\infty(\Omega, \mu) \rtimes G$ is a factor.

Krieger showed that in the case of a Krieger factor $M = L^\infty(\Omega, \mu) \rtimes G$, the invariant $S(M)$ agrees with the so-called “asymptotic ratio set” of the group G of nonsingular transformations, which is computable purely in terms of the Radon–Nikodym derivatives $d(\mu \circ T_t)/d\mu$. Using this ratio set description, it is not hard to see that the Krieger factor M given by the infinite product Ω

- is a factor of type III_λ if $\nu_n = 2$ and $\mu_n\{0\} = \lambda/(1 + \lambda)$ for all n ;
- is a factor of type III_1 if $\nu_n = 2$ and $\mu_{2n}\{0\} = \lambda/(1 + \lambda)$, $\mu_{2n+1}\{0\} = \kappa/(1 + \kappa)$, for all n , provided that $\{\lambda, \kappa\}$ generates a dense multiplicative subgroup of \mathbb{R}_+^\times ;
- can be of type III_0 .

Among all factors, Connes identified one tractable class – the so-called injective factors – which are ubiquitous and amenable to classification. To start

with, he established the equivalence of several (seemingly quite disparate) requirements on a von Neumann algebra $M \subset \mathcal{L}(\mathcal{H})$ – ranging from injectivity (meaning the existence of a projection of norm 1 from $\mathcal{L}(\mathcal{H})$ onto M) to “approximate finite dimensionality” (meaning $M = (\cup_n A_n)''$ for some increasing sequence $A_1 \subset A_2 \subset \dots \subset A_n \subset \dots$ of finite-dimensional $*$ -subalgebras). In the same paper, Connes (1976) essentially finished the complete classification of injective factors. Only the injective III_1 factor withstood his onslaught; but eventually even it had to surrender to the technical virtuosity of Haagerup (1987) a few years later!

In the language we have developed thus far, the classification of injective factors may be summarized as follows:

- Every injective factor is isomorphic to a Krieger factor.
- Up to isomorphism, there is a unique injective factor of each type with the solitary exception of III_0 .
- Injective factors of type III_0 are classified (up to isomorphism) by an invariant of an ergodic-theoretic nature called the “flow of weights”; unfortunately, coming up with a crisp description of this invariant, which is simultaneously accessible to the nonexpert and is consistent with the stipulated size of this survey, is beyond the scope of this author.

The interested reader is invited to browse through one of the books (Connes 1994, Sunder 1986, Dixmier 1981) for further details; the third book is the oldest (a classic but the language has changed a bit since it was written), the second is more recent (but quite sketchy in many places), and the first is clearly the best choice (if one has the time to read it carefully and digest it). Alternatively, the interested reader might want to browse through the encyclopediac treatments (Kadison and Ringrose) or (Takesaki).

See also: Algebraic Approach to Quantum Field Theory; Bicrossproduct Hopf Algebras and Noncommutative Spacetime; Braided and Modular Tensor Categories; C^* -Algebras and Their Classification; Ergodic Theory; Finite-Type Invariants; Hopf Algebra Structure of

Renormalizable Quantum Field Theory; Hopf Algebras and q -Deformation Quantum Groups; The Jones Polynomial; Knot Theory and Physics; Noncommutative Geometry and the Standard Model; Noncommutative Tori, Yang–Mills and String Theory; Positive Maps on C^* -Algebras; Quantum 3-Manifold Invariants; Quantum Entropy; Tomita–Takesaki Modular Theory; von Neumann Algebras: Subfactor Theory.

Further Reading

Connes A (1973) Une classification des facteurs de type III. *Annales Scientifiques de l'Ecole Normale Supérieure* 6: 133–252.

Connes A (1976) Classification of injective factors. *Annals of Mathematics* 104: 73–115.

Connes A (1994) *Non-commutative Geometry*. San Diego: Academic Press.

Dixmier J (1981) *von Neumann Algebras*. Amsterdam: North Holland.

Haagerup U (1987) Connes’ bicentralizer problem and uniqueness of the injective factor of type III_1 . *Acta Mathematica* 158: 95–148.

Kadison RV and Ringrose JR (1983/1986) *Fundamentals of the Theory of Operator Algebras*, vol. I–IV. New York: Academic Press.

Krieger W (1970) On the Araki–Woods asymptotic ratio set and non-singular transformations of a measure space. In: *Contributions to Ergodic Theory and Probability*, pp. 158–177. Lecture Notes in Mathematics, vol. 160, Springer-Verlag.

Krieger W (1976) On ergodic flows and the isomorphism of factors. *Mathematische Annalen* 223: 19–70.

Murray FJ and von Neumann J (1936) Rings of operators. *Annals of Mathematics* 37: 116–229.

Murray FJ and von Neumann J (1937) On rings of operators, II. *Transactions of the American Mathematical Society* 41: 208–248.

Murray FJ and von Neumann J (1943) On rings of operators, IV. *Annals of Mathematics* 44: 716–808.

Sakai S (1983) *C^* -algebras and W^* -algebras*. Berlin–New York: Springer-Verlag.

Sunder VS (1986) *An Invitation to von Neumann Algebras*. New York: Springer-Verlag.

Takesaki M (1970) *Tomita’s Theory of Modular Hilbert Algebras and its Applications*, Lecture Notes in Mathematics 128, Berlin–Heidelberg–New York: Springer-Verlag.

Takesaki M (1979/2003) *Theory of Operator Algebras*, vols. I–III. Heidelberg: Springer Verlag.

von Neumann J (1936) On rings of operators, III. *Annals of Mathematics* 37: 111–115.

von Neumann J (1949) On rings of operators. Reduction theory. *Annals of Mathematics*, 50: 401–485.

von Neumann Algebras: Subfactor Theory

Y Kawahigashi, University of Tokyo, Tokyo, Japan

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Subfactor theory was initiated by Jones (1983) and has experienced rapid progress beyond the framework of operator algebras. Here we start with a basic introduction in this section.

A factor is a von Neumann algebra with a trivial center. A von Neumann algebra M is an algebra of bounded linear operators on a Hilbert space H , which contains the identity operator and is closed under the $*$ -operation and weak operator topology, and its center is the intersection of M and its commutant

$$M' = \{x \in B(H) | xy = yx \text{ for all } y \in M\}$$

where $B(H)$ denotes the set of all the bounded linear operators on H . (We are mostly interested in separable, infinite-dimensional Hilbert spaces. A von Neumann algebra is automatically closed also in the norm topology and thus it is also a C^* -algebra.) By definition, a factor M acts on a certain Hilbert space H , but we also consider its action on another Hilbert space K , that is, a σ -weakly continuous homomorphism preserving the $*$ -operation from M into $B(K)$. A subfactor is a factor N which is contained in another factor M and has the same identity. A factor is classified into types I_n ($n=1, 2, 3, \dots$), I_∞ , II_1 , II_∞ , and III. In most of the interesting studies of subfactors, the two factors are of both type II_1 or both type III. A factor M is said to be of type II_1 if it is infinite dimensional and has a finite trace $\text{tr}: M \rightarrow \mathbb{C}$. By definition, a finite trace tr is a linear functional on M satisfying $\text{tr}(1)=1$, $\text{tr}(xy)=\text{tr}(yx)$ for all $x, y \in M$, and $\text{tr}(x^*x) \geq 0$ for all $x \in M$. When a factor M , not isomorphic to \mathbb{C} , acts on a separable Hilbert space, it is of type III if and only if for any two nonzero projections $p, q \in M$, we have an operator $v \in M$ with $vv^*=p$ and $v^*v=q$. One obviously cannot have a trace on such a factor. (See Takesaki (2002, 2003) for a general theory on factors.)

Let M be a type II_1 factor acting on a Hilbert space H . We then have the coupling constant of Murray and von Neumann, which is denoted by $\dim_M H$ and belongs to $(0, \infty]$. This measures the relative dimension of H with respect to M . Note that the factor M acts on M itself by the left multiplication. We introduce an inner product on M by $(x, y) = \text{tr}(y^*x)$ and denote the completion by

$L^2(M)$. Then M acts on this Hilbert space and we have $\dim_M L^2(M) = 1$.

Let $N \subset M$ be a subfactor and suppose that both N and M are of type II_1 . (We then simply say that $N \subset M$ is a type II_1 subfactor.) Suppose that M acts on a Hilbert space H with $\dim_M H < \infty$. Then we define the Jones index of N in M by

$$[M : N] = \frac{\dim_N H}{\dim_M H} \in [1, \infty]$$

This number is independent of the choice of H , as long as $\dim_M H < \infty$, so we can take $H = L^2(M)$, then we have $[M : N] = \dim_N L^2(M)$. The equality $[M : N] = 1$ means $M = N$. The first major discovery of Jones (1983) is that the value of the Jones index is in the set

$$\{4 \cos^2(\pi/m) | m = 3, 4, 5, \dots\} \cup [4, \infty] \quad [1]$$

and all the values in this set are indeed realized.

Suppose we have a II_1 factor M and an action of an at most countable, discrete group G on M , that is, a homomorphism $\alpha: G \rightarrow \text{Aut}(M)$, where $\text{Aut}(M)$ is the automorphism group of M . Then we have a construction $M \rtimes_\alpha G$, called the crossed product. If α_g is not an inner automorphism of M for any $g \in G$ other than the identity element of G , then $M \rtimes_\alpha G$ is also a type II_1 factor. (An automorphism π of M is said to be inner if it is of the form $\pi(x) = uxu^*$ for some unitary operator $u \in M$.) The index of a subfactor $M \subset M \rtimes_\alpha G$ is the order of G , which can be infinite. If we have a subgroup H of G , then we obtain a subfactor $M \rtimes_\alpha H \subset M \rtimes_\alpha G$ and its index is given by the index $[G : H]$ of the subgroup H . This analogy to the index of a subgroup is the origin of the terminology of the Jones index for a subfactor. The Jones index is also analogous to the degree of an extension of a field. From the viewpoint of this analogy, subfactor theory can be regarded as a certain generalized analogue (or the “quantum” version) of the classical Galois theory for field extensions. (The direct analog of the classical Galois correspondence for subfactors was studied by Nakamura–Takeda in the early days, and Izumi–Longo–Popa gave the most general form.)

The tools Jones (1983) has introduced to study subfactors are as follows. Let $N \subset M$ be a subfactor of type II_1 with finite Jones index. We consider the actions of N, M on $L^2(M)$. The completion of N with respect to the inner product given by the trace gives $L^2(N)$, which is naturally regarded as a closed subspace of $L^2(M)$. Let e_N be the projection on $L^2(M)$ onto $L^2(N)$, which is called the Jones

projection. We define M_1 to be the von Neumann algebra generated by M and e_N on $L^2(M)$. This is again a type II₁ factor and denoted by M_1 . This construction is called the basic construction. We obtain $[M_1 : M] = [M : N]$. Repeat the same procedure for $M \subset M_1$ acting on $L^2(M_1)$ this time. In this way, we have an increasing sequence of type II₁ factors,

$$N \subset M \subset M_1 \subset M_2 \subset M_3 \subset \dots$$

which is called the Jones tower. We label the corresponding Jones projections as $e_1 = e_N$, $e_2 = e_M$, $e_3 = e_{M_1}$, \dots . We then have the following celebrated Jones relations:

$$\begin{aligned} e_j e_k &= e_k e_j, & \text{if } |j - k| > 1 \\ e_j e_{j+1} e_j &= \frac{1}{[M : N]} e_j \end{aligned} \quad [2]$$

Jones proved the above-mentioned restriction on the possible values of the Jones index using these relations. The realization of the index values below 4 in the set [1] by Jones also relies on these relations of the Jones projection. The basic construction is also possible for the other direction. That is, we can construct a subfactor $N_1 \subset N$ so that $N \subset M$ is the basic construction of $N_1 \subset N$. This is called the downward basic construction. This N_1 is not unique, but is unique up to an inner automorphism of N .

A subfactor $N \subset M$ is said to be irreducible if the relative commutant $N' \cap M$ is equal to \mathbb{C} . If a subfactor has Jones index less than 4, then it is automatically irreducible. The original realization of the Jones index values above 4 by Jones was through reducible subfactors. Popa proved that all the values above 4 are realized with irreducible subfactors. A factor is said to be hyperfinite if it has a dense subalgebra given as the union of increasing sequence of finite-dimensional \ast -algebras. If M is a hyperfinite type II₁ factor, then its subfactor is automatically also hyperfinite by a deep theorem of Connes. For hyperfinite, irreducible type II₁ subfactors, it is still an open problem to determine all the possible values of the Jones index.

For type II₁ factors $N \subset M \subset P$, the Jones index $[P : N]$ is equal to the product $[P : M][M : N]$. Thus for the Jones tower, we have $[M_k : N] = [M : N]^{k+1}$. In general, if a subfactor $N \subset M$ has a finite Jones index, then the relative commutant $N' \cap M$ is automatically finite dimensional. So, if we start with a type II₁ subfactor $N \subset M$ with finite Jones index, we have an increasing sequence of finite-dimensional algebras as follows:

$$N' \cap M \subset N' \cap M_1 \subset N' \cap M_2 \subset N' \cap M_3 \subset \dots \quad [3]$$

These finite-dimensional algebras are called higher relative commutants of $N \subset M$. We draw the Bratteli diagram for the higher relative commutants as follows. Consider $N' \cap M_k$ (with convention $M_{-1} = N$, $M_0 = M$), then it is a finite-dimensional \ast -algebra; thus, it is of the form $\bigoplus_j M_{n_j}(\mathbb{C})$, where we have only finitely many direct summands. We draw a dot for each summand. We similarly draw a dot for each summand in $\bigoplus_l M_{m_l}(\mathbb{C})$ for $N' \cap M_{k+1}$. Let ι be the inclusion map from $N' \cap M_k = \bigoplus_j M_{n_j}(\mathbb{C})$ to $N' \cap M_{k+1} = \bigoplus_l M_{m_l}(\mathbb{C})$ and p_l the identity of $M_{m_l}(\mathbb{C})$, which is a projection in $N' \cap M_{k+1}$. We denote by μ_{jl} the multiplicity of the embedding map $x \mapsto \iota(x)p_l$ from $M_{n_j}(\mathbb{C})$ to $M_{m_l}(\mathbb{C})$. Then we draw μ_{jl} edges from the j th dot for $M_{n_j}(\mathbb{C})$ to the l th dot for $M_{m_l}(\mathbb{C})$. We repeat this procedure for all k , and get a picture as in Figure 1, which is called the Bratteli diagram of the higher relative commutants of $N \subset M$.

It turns out that the edges connecting the k th and $(k+1)$ th steps of the Bratteli diagram consist of the reflection of those connecting the $(k-1)$ th and k th steps, and a (possibly empty) new part. The “new” parts taken altogether in the above Bratteli diagram constitute the principal graph of a subfactor $N \subset M$. In the example of Figure 1, the principal graph is the Dynkin diagram A_5 . In general, a principal graph can be finite or infinite. If it is finite, we say that a subfactor is of finite depth. If a subfactor has the Jones index less than 4, it is automatically of finite depth and the principal graph must be one of the A – D – E Dynkin diagrams.

Pimsner and Popa (1986) obtained the characterization of the Jones index value in terms of the Pimsner–Popa inequality for a conditional expectation. This can be used as a definition of the index for a subfactor of arbitrary type (and even for C^\ast -subalgebras). Kosaki obtained a definition of the index for type III subfactors based on works of Connes and Haagerup.

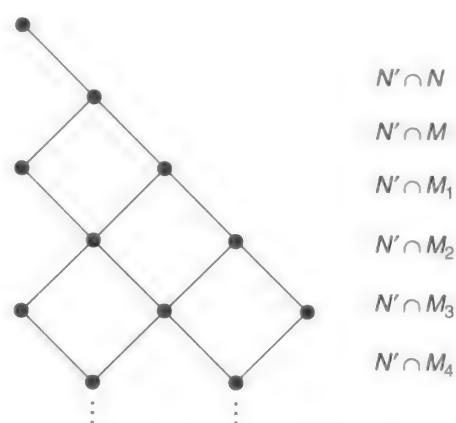


Figure 1 The Bratteli diagram of the higher relative commutants.

Analytic Classification Theory

If M is a hyperfinite type II_1 factor, then it is unique up to isomorphism. So any subfactor of such M is isomorphic to M itself. We next consider the classification problem of hyperfinite type II_1 subfactors. We say that a subfactor $N \subset M$ is isomorphic to $P \subset Q$ if we have an isomorphism of M onto Q which maps N onto P . The following tower of finite-dimensional algebras is a natural invariant for a type II_1 subfactor $N \subset M$ with finite Jones index and it is called the standard invariant for $N \subset M$:

$$\begin{array}{ccccccc} M' \cap M & \subset & M' \cap M_1 & \subset & M' \cap M_2 & \subset & \cdots \\ & \cap & & \cap & & \cap & \\ N' \cap M & \subset & N' \cap M_1 & \subset & N' \cap M_2 & \subset & \cdots \end{array} \tag{4}$$

Each square

$$\begin{array}{ccc} M' \cap M_k & \subset & M' \cap M_{k+1} \\ & \cap & \cap \\ N' \cap M_k & \subset & N' \cap M_{k+1} \end{array}$$

is a special combination of inclusions called a commuting square. Under a fairly general condition (called extremality of a subfactor, which automatically holds for an irreducible subfactor), the above sequence [4] is anti-isomorphic to the following sequence of finite-dimensional algebras, including the trace values:

$$\begin{array}{ccccccc} M' \cap M & \subset & N' \cap M & \subset & N'_1 \cap M & \subset & \cdots \\ & \cap & & \cap & & \cap & \\ M' \cap M_1 & \subset & N' \cap M_1 & \subset & N'_1 \cap M_1 & \subset & \cdots \end{array} \tag{5}$$

where $\cdots \subset N_3 \subset N_2 \subset N_1 \subset N \subset M$ is given by repeated downward basic constructions. So, if the closure of $\bigcup_j (N'_j \cap M_1)$ in the weak operator topology is equal to M_1 for an appropriate choice of N'_j 's, then the closure of $\bigcup_j (N'_j \cap M)$ is also M , and the isomorphism class of the subfactor $N \subset M$ is recovered from the standard invariant. In such a case, we say that a subfactor has a generating property, and then we have a complete classification of subfactors in terms of the standard invariant. Popa (1994) introduced a notion called strong amenability and proved that a subfactor of type II_1 is strongly amenable if and only if it has the generating property. This is the fundamental result in the classification of subfactors. A hyperfinite type II_1 subfactor with finite Jones index and finite depth is automatically strongly amenable, so such a subfactor is covered by this classification theorem of Popa. Popa also has a similar result for subfactors of type III .

Constructions and Combinatorial Classification

As mentioned in the above section, Jones constructed hyperfinite type II_1 subfactors for all possible index values below 4. They have the Dynkin diagrams A_n as the principal graphs. It has been an important problem to construct new subfactors since then. Using the Hecke algebras, Wenzl constructed a series of subfactors with index values $\sin^2(N\pi/k)/\sin^2(\pi/k)$ with $N=2,3,4,\dots$, where the series for $N=2$ coincide with the ones constructed by Jones. Wenzl's dimension estimate in this work for the relative commutant has been an important tool to study subfactors. It was soon noticed that the subfactors of Jones and Wenzl are related to the quantum groups $U_q(\mathfrak{sl}_N)$ of Drinfel'd-Jimbo, at the value of the deformation parameter q at $\exp(\pi i/k)$. Constructions of subfactors from other quantum groups have been given by Wenzl.

Ocneanu (1988) has introduced a notion of a paragroup and characterized the higher relative commutants arising from a type II_1 subfactor with finite Jones index and finite depth as a paragroup. If we start with a subfactor $N \subset N \rtimes_{\sigma} G$ for a finite group G , the corresponding paragroup contains complete information on the group G and its representations. In this sense, a paragroup is a generalization of a (finite) group. The basic idea is to regard the bimodule ${}_N L^2(M)_M$ as an analog of the fundamental representation of a compact Lie group and make finite relative tensor products

$$\cdots \otimes_N L^2(M) \otimes_M L^2(M) \otimes_N L^2(M) \otimes_M \cdots$$

Then one makes an irreducible decomposition and studies various intertwiners arising from these irreducible bimodules. In this way, we obtain a certain combinatorial object and it is called a paragroup. The vertices of the principal graphs correspond to irreducible bimodules and the edges correspond to basis vectors in the intertwiner spaces. Note that by Popa's theorem explained in the previous section, a classification of subfactors of a hyperfinite type II_1 factor with finite Jones index and finite depth is reduced to one of paragroups.

Using this theory of paragroups, Ocneanu has found that the Dynkin diagrams A_n, D_{2n}, E_6 , and E_8 are realized as principal graphs of subfactors, but D_{2n+1} and E_7 are not. Furthermore, each of the graphs A_n and D_{2n} has unique realization and each of E_6, E_8 has two realizations. At the index value 4, the principal graph must be one of the extended Dynkin diagrams, $A_{2n-1}^{(1)}, D_n^{(1)}, E_6^{(1)}, E_7^{(1)}, E_8^{(1)}, A_{\infty}, A_{\infty, \infty}$, and D_{∞} , and all are realized. (The last three correspond to subfactors of infinite depth.)

See Evans and Kawahigashi (1998) and Goodman *et al.* (1989) for these constructions and classifications. Evans-Kawahigashi and Xu studied the orbifold construction of subfactors applied to the Hecke algebra subfactors of Wenzl.

In a theory of integrable lattice models, we have squares with labeled edges, and we assign complex numbers to them. A paragroup has much formal similarity to such a lattice model, and the paragroups of subfactors of Jones and Wenzl correspond to the lattice models of Andrews-Baxter-Forrester.

Goodman-de la Harpe-Jones have another construction of subfactors from the Dynkin diagrams, and for E_6 this gives a hyperfinite type II_1 subfactor with Jones index $3 + \sqrt{3}$ and finite depth. Haagerup has made a combinatorial study on type II_1 subfactors with Jones index values between 4 and $3 + \sqrt{3}$ and obtained a list of candidates of possible higher relative commutants. Haagerup himself showed one in the list with Jones index $(5 + \sqrt{13})/2$ is indeed realized. Asaeda-Haagerup showed that another in the list having the Jones index $(5 + \sqrt{17})/2$ is also realized. These two examples are still among the most mysterious examples of subfactors today and do not seem to arise from other constructions using quantum groups or conformal field theory. Izumi has another construction of a subfactor with the Jones index $(7 + \sqrt{29})/2$ using an endomorphism of the Cuntz algebra.

Popa has obtained a complete characterization of higher relative commutants including the case of infinite depth, and axiomatized the higher relative commutant as the standard λ -lattices. Xu has constructed standard λ -lattices, hence subfactors, from quantum groups. This realization of Popa of a given standard λ -lattice produces a nonhyperfinite type II_1 subfactor. Popa-Shlyakhtenko later showed that any standard λ -lattice is realized for a subfactor of a single type II_1 factor, a group II_1 factor arising from the free group F_∞ having countably many generators, which is not hyperfinite.

Jones (1999) has introduced a combinatorial characterization of standard λ -lattices as planar algebras. This approach uses planar operads based on tangles and provides a new viewpoint on the structure of higher relative commutants. More studies on planar algebras have been done by Bisch-Jones.

Topological Invariants in Three Dimensions and Tensor Categories

Through the relations of the Jones projections, Jones (1985) discovered the Jones polynomial as an

invariant for links. This was the beginning of series of entirely new theories in three-dimensional topology. The Jones polynomial was quickly generalized to the two-variable HOMFLY polynomial by Hoste, Ocneanu, Millet, Freyd, Lickorish, and Yetter.

A three-dimensional topological quantum field theory (TQFT₃) assigns a complex number to each closed oriented 3-manifold and a finite dimensional vector space to each closed oriented surface. Furthermore, to each compact oriented 3-manifold with boundary, it assigns a vector in the vector space corresponding to its boundary. Turaev-Viro have constructed TQFT₃ from combinatorial data called quantum $6j$ -symbols arising from quantum groups. Ocneanu has found that a subfactor of finite index and finite depth also produces quantum $6j$ -symbols, which give rise to a TQFT₃ generalizing the Turaev-Viro construction. See Evans and Kawahigashi (1998) for this construction. Reshetikhin-Turaev have another construction of TQFT₃ from a modular tensor category, which is a braided tensor category with nondegenerate braiding. Ocneanu has found a subfactor version of the quantum double construction which produces a modular tensor category from a type II_1 subfactor of finite index and finite depth. From a type II_1 subfactor of finite index and finite depth, we can apply Ocneanu's generalization of the Turaev-Viro construction on one hand, and also the Reshetikhin-Turaev construction to the modular tensor category arising from the quantum double construction of Ocneanu. The resulting two TQFT₃s are shown to be equal by Kawahigashi-Sato-Wakui. Concrete computations of these topological invariants have been made by Sato-Wakui based on Izumi's work. Turaev and Wenzl have other constructions of TQFT₃ and modular tensor categories.

Algebraic Quantum Field Theory

An operator algebraic approach to quantum field theory is called algebraic quantum field theory and the standard reference is Haag (1996). In this approach, instead of quantum fields which are operator-valued distributions, we consider a family $\{A(O)\}$ of von Neumann algebras parametrized by spacetime regions O in a Minkowski space. Each $A(O)$ is meant to be generated by self-adjoint operators which are observables in O . We axiomatize such a family of von Neumann algebras and call one a local net of von Neumann algebras. It is enough to take O of a special form, called a double cone. The name "local" comes from the locality axiom which is a mathematical expression of the Einstein causality on a Minkowski space. The

Poincaré group is used as the spacetime symmetry of the Minkowski space. Doplicher *et al.* (1971, 1974) have introduced a representation theory of a local net A of von Neumann algebras and found that a “physically nice” representation is realized as an endomorphism of a one von Neumann algebra $A(O)$ for some fixed O . They have a notion of a statistical dimension for such a representation and it is an integer (or infinite) if the spacetime dimension is larger than 2. Longo (1989, 1990) has shown that this statistical dimension of a representation is equal to the square root of the index $[A(O) : \lambda(A(O))]$, where λ is the corresponding endomorphism of $A(O)$ to the representation. The relation between algebraic quantum field theory and subfactor theory has been found in this way. Longo (1989, 1990) has also started a theory of canonical endomorphisms for a subfactor and Izumi has further studied it. Longo has later obtained a characterization when an endomorphism of a factor becomes a canonical endomorphism by introducing a Q -system.

Recently, conformal field theory has attracted much attention. An approach based on algebraic quantum field theory describes a conformal field theory with a local net of von Neumann algebras on a two-dimensional Minkowski space with diffeomorphism group as the spacetime symmetry. We can restrict such a theory into a tensor product of two theories on the circle, the compactified one-dimensional Euclidean space. Each theory on the circle is called a chiral conformal field theory and described by a local conformal net of von Neumann algebras, which is a family of von Neumann algebras parametrized by intervals on the circle. The name “conformal” comes from the fact that we use the orientation preserving diffeomorphism group on the circle as the symmetry group of the space. For a local conformal net A of von Neumann algebras on the circle with natural irreducibility assumption, each von Neumann algebra $A(I)$ is automatically a type III factor. The Doplicher–Haag–Roberts theory works in this setting after an appropriate adaptation as in Fredenhagen *et al.* (1989) and each representation of a local conformal net of von Neumann algebras is realized by an endomorphism of $A(I)$, where I is an arbitrarily fixed interval on the circle. (Here we do not need an assumption that a representation is “physically nice” since it now automatically holds.) Now the representations give a braided tensor category.

Buchholz–Mack–Todorov constructed examples of local conformal nets of von Neumann algebras on the circle using the $U(1)$ -current algebra. Wassermann (1998) has constructed more examples using positive energy representations of the loop groups $LSU(N)$

and computed their representation theory, and his construction has been extended to other Lie groups by Toledano Laredo and others. For the local conformal net A of von Neumann algebras on the circle arising from $LSU(N)$, we take an endomorphism λ of $A(I)$ arising from a representation of the local conformal net, then we have a subfactor $\lambda(A(I)) \subset A(I)$. This is isomorphic to the type II_1 subfactor constructed by Jones and Wenzl tensored with a common type III factor.

Longo–Rehren (1995) started the study of a local net of subfactors, $A(I) \subset B(I)$. They have defined a certain induction procedure which gives a representation of the larger local conformal net B from that of A . This procedure is today called α -induction. Xu has studied this procedure and found several basic properties. In the cases of local conformal nets of subfactors arising from conformal embeddings, he has found a simple construction of subfactors with principal graphs E_6 and E_8 using α -induction. In the context of subfactor theory, α -induction has been further studied by Böckenhauer–Evans–Kawahigashi, together with graphical methods of Ocneanu on the Dynkin diagrams. More detailed studies on local conformal nets of factors on the circle have been pursued partly using various techniques of subfactor theory, including classification of local conformal nets of von Neumann algebras on the circle with central charge less than 1 by Kawahigashi–Longo.

See also: Algebraic Approach to Quantum Field Theory; Braided and Modular Tensor Categories; C^* -Algebras and Their Classification; Hopf Algebras and q -Deformation Quantum Groups; The Jones Polynomial; Quantum 3-Manifold Invariants; Quantum Entropy; von Neumann Algebras: Introduction, Modular Theory, and Classification Theory; Yang–Baxter Equations.

Further Reading

- Doplicher S, Haag R, and Roberts JE (1971) Local observables and particle statistics, I. *Communications in Mathematical Physics* 23: 199–230.
- Doplicher S, Haag R, and Robert JE (1974) Local observables and particle statistics, II. *Communications in Mathematical Physics* 35: 49–85.
- Evans DE and Kawahigashi Y (1998) *Quantum Symmetries on Operator Algebras*. Oxford: Oxford University Press.
- Fredenhagen K, Rehren K-H, and Schroer B (1989) Superselection sectors with braid group statistics and exchange algebras. *Communications in Mathematical Physics* 125: 201–226.
- Goodman F, de la Harpe P, and Jones VFR (1989) *Coxeter Graphs and Towers of Algebras*, vol. 14. Berlin: MSRI Publications, Springer.
- Haag R (1996) *Local Quantum Physics*. Berlin: Springer.
- Jones VFR (1983) Index for subfactors. *Inventiones Mathematicae* 72: 1–25.

Jones VFR (1985) A polynomial invariant for knots via von Neumann algebras. *Bulletin of the American Mathematical Society* 12: 103–112.

Jones VFR (1999) Planar algebras, I, math.QA/9909027.

Longo R (1989) Index of subfactors and statistics of quantum fields I. *Communications in Mathematical Physics* 126: 217–247.

Longo R (1990) Index of subfactors and statistics of quantum fields II. *Communications in Mathematical Physics* 130: 285–309.

Longo R and Rehren K-H (1995) Nets of subfactors. *Reviews in Mathematical Physics* 7: 567–597.

Ocneanu A (1988) Quantized group, string algebras and Galois theory for algebras. In: Evans DE and Takesaki M (eds.) *Operator Algebras and Applications*, vol. 2 (Warwick 1987),

London Mathematical Society Lecture Note Series, vol. 36, pp. 119–172. Cambridge: Cambridge University Press.

Pimsner M and Popa S (1986) Entropy and index for subfactors. *Annales Scientifiques de l'Ecole Normale* 19: 57–106.

Popa S (1994) Classification of amenable subfactors of type II. *Acta Mathematica* 172: 163–255.

Popa S (1995) An axiomatization of the lattice of higher relative commutants of a subfactor. *Inventiones Mathematicae* 120: 427–445.

Takesaki M (2002, 2003) *Theory of Operator Algebras I, II, III*. Berlin: Springer.

Wassermann A (1998) Operator algebras and conformal field theory III: Fusion of positive energy representations of SU(N) using bounded operators. *Inventiones Mathematicae* 133: 467–538.

Vortex Dynamics

M Nitsche, University of New Mexico, Albuquerque, NM, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

A vortex is commonly associated with the rotating motion of fluid around a common centerline. It is defined by the vorticity in the fluid, which measures the rate of local fluid rotation. Typically, the fluid circulates around the vortex, the speed increases as the vortex is approached and the pressure decreases. Vortices arise in nature and technology applications in a large range of sizes, as illustrated by the examples given in Table 1. The next section presents some of the mathematical background necessary to understand vortex formation and evolution. Next, some sample flows are described, including important instabilities and reconnection processes. Finally, some of the numerical methods used to simulate these flows are presented.

Background

Let D be a region in three-dimensional (3D) space containing a fluid, and let $\mathbf{x} = (x, y, z)^T$ be a point in

D . The fluid motion is described by its velocity $\mathbf{u}(\mathbf{x}, t) = u(\mathbf{x}, t)\mathbf{i} + v(\mathbf{x}, t)\mathbf{j} + w(\mathbf{x}, t)\mathbf{k}$, and depends on the fluid density $\rho(\mathbf{x}, t)$, temperature $T(\mathbf{x}, t)$, gravitational field \mathbf{g} , and other external forces possibly acting on it. The fluid vorticity is defined by $\boldsymbol{\omega} = \nabla \times \mathbf{u}$. The vorticity measures the local fluid rotation about an axis, as can be seen by expanding the velocity near $\mathbf{x} = \mathbf{x}_0$,

$$\mathbf{u}(\mathbf{x}) = \mathbf{u}(\mathbf{x}_0) + D(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}\boldsymbol{\omega}(\mathbf{x}_0) \times (\mathbf{x} - \mathbf{x}_0) + O(|\mathbf{x} - \mathbf{x}_0|^2)$$
 [1]

where

$$D(\mathbf{x}_0) = \frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^T), \quad \nabla \mathbf{u} = \begin{bmatrix} u_x & u_y & u_z \\ v_x & v_y & v_z \\ w_x & w_y & w_z \end{bmatrix}$$
 [2]

The first term $\mathbf{u}(\mathbf{x}_0)$ corresponds to translation: all fluid particles move with constant velocity $\mathbf{u}(\mathbf{x}_0)$. The second term $D(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$ corresponds to a strain field in the three directions of the eigenvectors of the symmetric matrix D . If the eigenvalue corresponding to a given eigenvector is positive, the fluid is stretched in that direction, if it is negative, the fluid is compressed. Note that, in incompressible flow, $\nabla \cdot \mathbf{u} = 0$, so the sum of the eigenvalues of D equals zero. Thus, at least one eigenvalue is positive and one negative. If the third eigenvalue is positive, fluid particles move towards sheets (Figure 1a). If the third eigenvalue is negative, fluid particles move towards tubes (Figure 1b). The last term in eqn [1], $(1/2)\boldsymbol{\omega}(\mathbf{x}_0) \times (\mathbf{x} - \mathbf{x}_0)$, corresponds to a rotation: near a point with $\boldsymbol{\omega}(\mathbf{x}_0) \neq 0$, the fluid rotates with angular velocity $|\boldsymbol{\omega}|/2$ in a plane normal to the vorticity vector $\boldsymbol{\omega}$. Fluid for which $\boldsymbol{\omega} = 0$ is said to be irrotational.

A vortex line is an integral curve of the vorticity. For incompressible flow, $\nabla \cdot \boldsymbol{\omega} = \nabla \cdot (\nabla \times \mathbf{u}) = 0$, which implies that vortex lines cannot end in the

Table 1 Sample vortices and typical sizes

Vortex	Diameter
Superfluid vortices	10 ⁻⁸ cm (= 1 Å)
Trailing vortex of Boeing 727	1–2 m
Dust devils	1–10 m
Tornadoes	10–500 m
Hurricanes	100–2000 km
Jupiter's Red Spot	25 000 km
Spiral galaxies	Thousands of light years

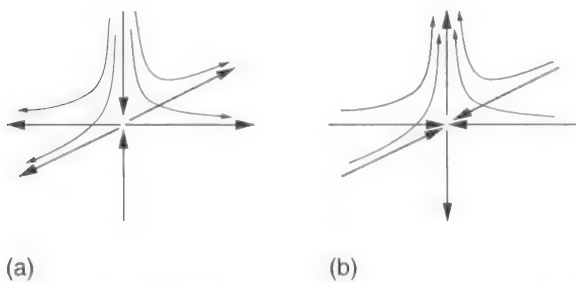


Figure 1 Strain field: (a) two positive eigenvalues, sheet formation and (b) one positive eigenvalue, tube formation.

interior of the flow, but must either form a closed loop or start and end at a bounding surface. In 2D flow, $\mathbf{u} = u\mathbf{i} + v\mathbf{j}$ and the vorticity is $\boldsymbol{\omega} = \omega\mathbf{k}$, where ω is the scalar vorticity. Thus, in 2D, the vorticity points in the z -direction and the vortex lines are straight lines normal to the x - y plane. A vortex tube is a bundle of vortex lines. The strength of a vortex tube is defined as the circulation $\int_C \mathbf{u} \cdot d\mathbf{s}$ about a curve C enclosing the tube. By Stokes' theorem,

$$\int_C \mathbf{u} \cdot d\mathbf{s} = \iint_A \boldsymbol{\omega} \cdot \mathbf{n} dS \tag{3}$$

and thus the circulation can also be interpreted as the flux of vorticity through a cross section of the tube. In inviscid incompressible flow of constant density, Helmholtz' theorem states that the tube strength is independent of the curve C , and is therefore a well-defined quantity, and Kelvin's theorem states that a tube's strength remains constant in time. A vortex filament is an idealization in which a tube is represented by a single vortex line of nonzero strength.

The evolution equation for the fluid vorticity, as derived from the Navier–Stokes equations, is

$$\frac{d\boldsymbol{\omega}}{dt} = \boldsymbol{\omega} \cdot \nabla \mathbf{u} + \nu \Delta \boldsymbol{\omega} \tag{4}$$

where $d/dt = \partial/\partial t + \mathbf{u} \cdot \nabla$ is the total time derivative. Equation [4] states that the vorticity is transported by the fluid velocity (first term), stretched by the fluid velocity gradient (second term), and diffused by viscosity ν (last term). These equations are usually nondimensionalized and written in terms of the Reynolds number, a dimensionless quantity inversely proportional to viscosity.

To understand high Reynolds number flow it is of interest to study the inviscid Euler equations. The corresponding vorticity evolution equation in 2D is

$$\frac{d\omega}{dt} = 0 \tag{5}$$

which states that 2D vortex filaments in inviscid flow move with the fluid velocity. Furthermore, in

incompressible flow, the fluid velocity is determined by the vorticity, up to an irrotational far-field component \mathbf{u}_∞ , through the Biot–Savart law,

$$\mathbf{u}(\mathbf{x}) = -\frac{1}{4\pi} \int \frac{(\mathbf{x} - \mathbf{x}') \times \boldsymbol{\omega}(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|^3} d\mathbf{x}' + \mathbf{u}_\infty \tag{6}$$

In planar 2D flow, eqn [6] reduces to

$$\mathbf{u}(\mathbf{x}) = K_{2D}\boldsymbol{\omega}, \quad K^*_{2D}(\mathbf{x}) = \frac{1}{2\pi} \frac{-y\mathbf{i} + x\mathbf{j}}{|\mathbf{x}|^2} \tag{7}$$

where $\omega(\mathbf{x})$ is the scalar vorticity. Equations [4], [5] and [6], [7] are the basis of the numerical methods discussed later in this article.

A vortex is typically defined by a region in the fluid of concentrated vorticity. A simple model is a point vortex in 2D flow, which corresponds to a straight vortex filament of unit circulation. The associated scalar vorticity is a δ -function in the plane, and the induced velocity is obtained from the Biot–Savart law. For a point vortex at the origin, this reduces to the radial velocity field $\mathbf{u}(\mathbf{x}) = K^*_{2D}\delta = K_{2D}(\mathbf{x})$. Corresponding particle trajectories are shown in **Figure 2a**. The particle speed $|\mathbf{u}| = 1/(2\pi r)$ increases unboundedly as the vortex center is approached, and vanishes as $r \rightarrow \infty$ (**Figure 2b**). In general, the far-field velocity of a concentrated vortex behaves similarly to the one of a point vortex, with speeds decaying as $1/r$. Near the vortex center, the velocity typically increases in magnitude and, as a result, the fluid pressure decreases (Bernoulli's theorem). A vortex of arbitrary shape can be approximated by a sum of point vortices (in 2D) or vortex filaments (in 3D), as is often done for simulation purposes.

Vorticity can be generated by a variety of mechanisms. For example, vorticity can be generated by density gradients, which in turn are induced by spatial temperature variations. This mechanism explains the formation of warm-air vortices when a layer of hot air is trapped underneath cooler air. Vorticity is also generated near solid walls in the form of boundary layers caused by viscosity. To illustrate, imagine

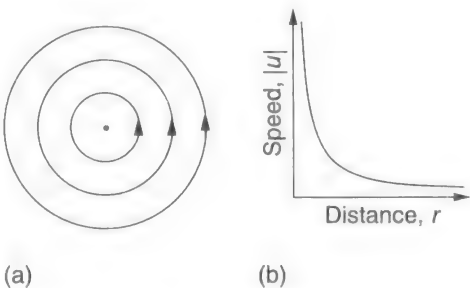


Figure 2 Flow induced by a point vortex: (a) streamlines and (b) speed $|\mathbf{u}|$ vs. distance r .

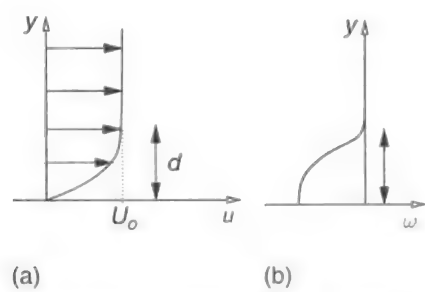


Figure 3 Velocity and vorticity in boundary layer near a flat wall.

horizontal flow with speed U_o moving past a solid wall at rest (Figure 3a). Since in viscous flow the fluid sticks to the wall (the no-slip boundary condition), the fluid velocity at the wall is zero. As a result, there is a thin layer near the wall in which the horizontal velocity varies greatly while the vertical velocity gradients are small, yielding large negative vorticity values $\omega = v_x - u_y$ (Figure 3b). Similarity solutions to the approximating Prandtl boundary-layer equations show that the boundary-layer thickness d grows proportional to \sqrt{t} , where t measures the time from the beginning of the motion. Boundary layers can separate from the wall at corners or regions of high curvature and move into the fluid interior, as illustrated in several of the following examples.

Sample Vortex Flows

Shear Layers

A shear layer is a thin region of concentrated vorticity across which the tangential velocity component varies greatly. An example is the constant-vorticity layer given by parallel 2D flow $u(x, y) = U(y)$, $v(x, y) = 0$, where U is as shown in Figure 4a. In this case, the velocity is constant outside the layer and linear inside. The vorticity $\omega = -U'(y)$ is zero outside the layer and constant

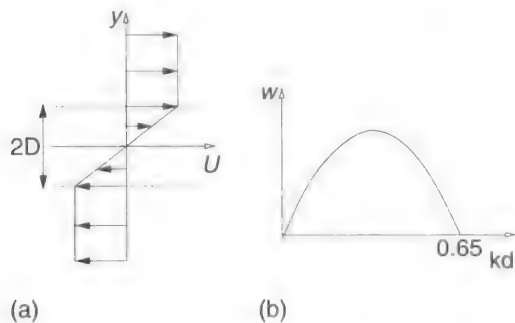


Figure 4 Shear layer: (a) velocity profile and (b) dispersion relation.

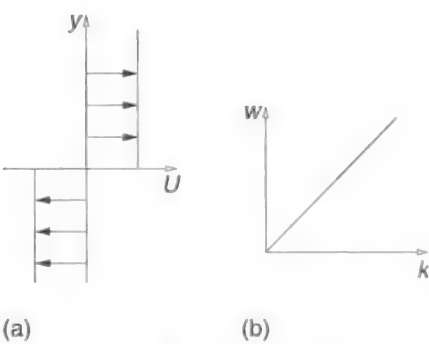


Figure 5 Vortex sheet: (a) velocity profile and (b) dispersion relation.

inside. Shear layers occur naturally in the ocean or atmosphere when regions of distinct temperature or density meet. To illustrate this scenario, consider a tank containing two horizontal layers of fluids of different densities, one on top of the other. If the tank is tilted, the heavier bottom fluid moves downstream, and the lighter one moves upstream, creating a shear layer.

Flat shear layers are unstable to perturbations: they do not remain flat but roll up into a sequence of vortices. This is the Kelvin–Helmholtz instability, which can be deduced analytically using linear stability analysis. One shows that in a periodically perturbed flat shear layer, the amplitude of a perturbation with wave number k will initially grow exponentially in time as e^{wt} , where $w = w(k)$ is the dispersion relation, leading to instability. The wave number of largest growth depends on the layer thickness. This is illustrated in Figure 4b, which plots $w(k)$ for a constant-vorticity layer of thickness $2d$. The wave number of maximal growth is proportional to $1/d$.

A vortex sheet is a model for a shear layer. The layer is approximated by a surface of zero thickness across which the tangential velocity is discontinuous, as illustrated in Figure 5a. In this case, the dispersion relation reduces to $w(k) = \pm k$. That is, for each wave number k there is a growing and a decaying mode, and the growing mode grows faster the higher the wave number is, as shown in Figure 5b. The vortex sheet arises from a constant vorticity shear layer as the thickness $d \rightarrow 0$ and the vorticity $\omega \rightarrow \infty$ in such a way that the product ωd remains constant. Figure 6 shows the roll-up of a periodically perturbed vortex sheet due to the



Figure 6 Computation of vortex sheet roll-up.

Kelvin–Helmholtz instability, computed using one of the methods described in the next section.

Aircraft Trailing Vortices

One can often observe trailing vortices that shed from the wings of a flying aircraft (also called contrails). These vortices are formed because the wing develops lift. The pressure on the top of the wing is lower than on bottom, causing air to move around the edge of the wing from the bottom surface to the top. The boundary layer on the wing separates as a shear layer that rolls up into a vortex attached to the tip of the wing (Figure 7). Since the velocity inside the vortex is high, the pressure is correspondingly low and causes water vapor in the air to condense, forming water droplets that visualize the vortices. The vortex strength increases with increasing lift, and is particularly strong in high-lift conditions such as take-off and landing. Since lift is proportional to weight, it also increases with the size of the airplane. Vortices of large planes are strong enough to flip a small one if it gets too close. Trailing vortices are the principal reason for the time delay between take-off and landing and are still a serious issue for crowded urban airports.

The trailing vortices can be modeled by a pair of counter-rotating vortex lines (Figure 8a). Two parallel vortex lines of opposite strength induce a downward motion on each other, similar to two point vortices, the zero-core limit. Two point vortices of strength $\pm\Gamma$ at a distance $2d$ from each other translate with self-induced velocity (Figure 9):

$$U = \frac{\Gamma}{4\pi d} \tag{8}$$

As a result trailing vortices near takeoff hit the ground as a strong downwash air current.

Vortex decay results generally from the development of instabilities. Two parallel vortex tubes are subject to the long-wavelength Crow instability. Triggered by turbulence in the surrounding air, or by local variations in air temperature or density, the vortices develop symmetric sinusoidal perturbations with long wavelength, of the same order as the vortex separation (Figure 8b). As the perturbations grow to finite amplitude, the tubes reconnect and produce a sequence of vortex rings. Note that the

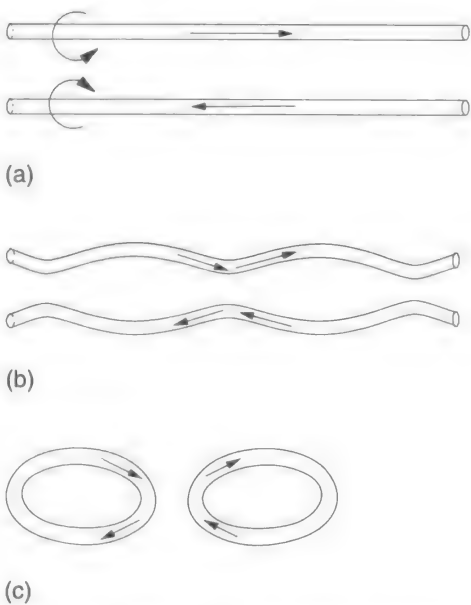


Figure 8 Sketch. Onset of Crow instability in a pair of vortex lines and ensuing reconnection.

two-dimensional schematic in Figure 8c does not convey the three-dimensional structure of the rings. The reconnection process destroys the initial wake structure more rapidly than viscous decay of the individual filaments.

Of much interest is the study of how to accelerate the vortex decay. High-aspect-ratio vortices are subject to a shorter-wavelength elliptic instability, which leads to earlier destruction. However, such vortices are not realistic in current aircraft wakes. Wing designs have been proposed in which more than two trailing vortices form which interact strongly and lead to faster decay. Other interesting aspects are the effect of ambient turbulence and vortex breakdown. Breakdown refers to a disturbance in the vortex core in which it quickly, within an axial distance of few core diameters, develops a region of reversed flow and loses its laminar behavior.

Unlike the counter-rotating vortices discussed so far, two equally signed vortices rotate under their self-induced velocity about a common axis. If the separation distance between them is too small, two equally signed patches merge into one. Vortex merging occurs in two- or three-dimensional flows, as opposed to vortex reconnection, which is a strictly three-dimensional phenomenon.

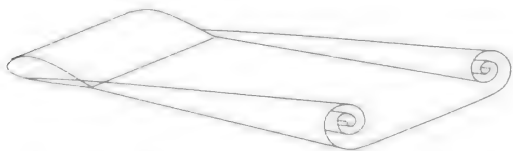


Figure 7 Sketch. Shear layer separation and roll-up into trailing vortices behind an airfoil.

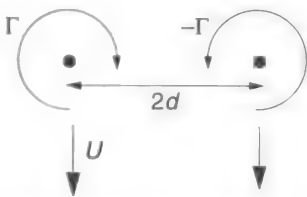


Figure 9 Self-induced downward motion of a vortex pair.

Vortex Rings

A vortex tube that forms a closed loop is called a vortex ring. Vortex rings can be formed by ejecting fluid from a circular opening, such as when a smoke ring is formed. The boundary layer wall vorticity separates at the opening as a cylindrical shear layer that rolls up at its edge into a ring (Figure 10). The vorticity is concentrated in a core, which may be thin or thick relative to the ring diameter. The limiting cases are an infinitely thin circular filament of nonzero circulation and the Hill’s vortex, in which the vorticity occupies all the interior of a sphere.

Just as a counter-rotating vortex pair, a ring translates under its self-induced velocity U in direction normal to the plane of the ring (Figure 11). However, unlike the vortex pair, the ring velocity depends significantly on its core thickness. For a ring with radius, circulation and core size, respectively, R, Γ, a , the self-induced velocity is

$$U \sim \frac{\Gamma}{4\pi R} \left(\log \frac{8R}{a} - \frac{1}{4} \right) \tag{9}$$

asymptotically as $a \rightarrow 0$. Thus, the translation velocity becomes unbounded for rings with decreasing core size. In reality, at some point viscosity takes over and spreads the core vorticity, slowing the ring down.

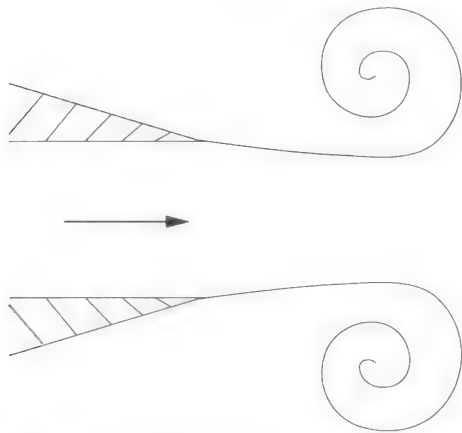


Figure 10 Vortex ring, formed by ejecting fluid from a circular tube.

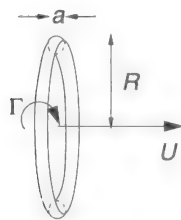


Figure 11 Self-induced motion of a vortex ring.

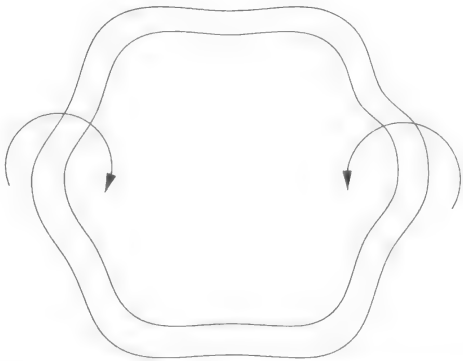


Figure 12 Sketch. Onset of azimuthal vortex ring instability.

Vortex rings of small cross section are subject to an azimuthal instability. Theory, experiment, and simulations show that if a ring is perturbed in the azimuthal direction, there exists a dominant wave number which is unstable and grows (Figure 12). The unstable wave number increases as the core size decreases, while its spatial amplification rate is almost independent of the core size.

Interesting dynamics are obtained when two or more rings interact. Two coaxial vortex rings of equally signed circulation move in the same direction and exhibit leap-frogging: the rear ring causes the front ring to grow in radius and the front ring causes the rear one to decrease. From eqn [9] it can be seen that the ring velocity is inversely proportional to its radius. Consequently, the front ring slows down and the rear ring speeds up, until the rear ring travels through the front ring. This process repeats itself and is known as leap-frogging. On the other hand, two coaxial vortex rings of oppositely signed circulation approach each other and grow in radius. Their cores contract in order to preserve volume, and their vorticity increases in order to preserve circulation. Under certain experimental conditions, the azimuthal instability develops, the resulting waves on opposite rings reconnect and a sequence of smaller rings form.

Vortices, Mixing, and Chaos

Mixing is important in many natural processes and technological applications. For example, mixing in shear flows and wakes is relevant to aeronautics and combustion, mixing and diffusion determine chemical reaction rates, and mixing of contaminants pollutes oceans and atmosphere. It is therefore important to understand and control mixing processes.

Efficient mixing of two fluids is obtained by efficient stretching and folding of material lines.

Stretching and folding in turn are the fingerprint of chaos; thus, mixing and chaos are intimately related. Mixing and associated chaotic fluid motion can be obtained by simple vortical motion. For example, two counter-rotating vortices subject to a periodic strain field oscillate in a regular fashion but induce chaos in a region of fluid moving with them. Similarly, two corotating vortices of equal strength that are turned on and off periodically so that one is on when the other is off, known as the blinking vortices, rotate around a common axis in a stepwise manner but induce chaos in nearby regions. On the other hand, if there are four or more vortices present, the vortex motion itself is generally chaotic. It should be noted that there are also nonchaotic equilibrium solutions of four or more vortices forming what is called a vortex crystal.

Information about chaotic particle motion is obtained by studying Poincaré sections, examining the associated stable and unstable manifolds, and investigating the existence of chaotic maps such as the horseshoe map.

Atmospheric Vortices

Atmospheric vortices are driven by temperature gradients, Earth's rotation (Coriolis force), spatial landscape variations, and instabilities. For example, temperature differences between the equator and the poles and Earth's rotation lead to large-scale vortices such as the trade winds (Hadley cell), the jet streams, and the polar vortex (Figure 13). Semi-annual temperature oscillations are responsible for the Indian monsoons. Daily oscillations cause land- and sea-breezes. Landscape variations can cause urban-rural wind flows and mountain-valley circulations.

Instabilities are often responsible for large cyclonic vortices. Barotropic instability results from large horizontal velocity gradients, and has been deemed responsible for disturbances over the Sahara region that occasionally intensify into

tropical cyclones. Baroclinic instability, which occurs when temperature advection is superposed on a velocity field, can lead to cyclonic vortices at the front between air of polar origin and that of tropical origin. The inertial or centrifugal instability occurs when air flows around high-pressure systems and the pressure gradient force is not large enough to balance the centripetal acceleration and the Coriolis effect.

Vortices also form on other planets with an atmosphere. On Mars, dust devils are quite common. They are ~ 10 – 50 times larger than the ones on Earth and can carry high-voltage electric fields caused by the rubbing of dust grains against each other. Jupiter's characteristic spots are extremely large storm vortices. The Great Red Spot is a vortex spanning twice the diameter of the Earth. Unlike the low-pressure terrestrial storms and hurricanes, the Great Red Spot is a high-pressure system that has been stable for more than 300 years. Other vortices on Jupiter decay and vanish, such as the White Ovals, three large anticyclones which merged into one within two years. Recent computer simulations predict that many of Jupiter's vortices will merge and disappear in the next decade. As a result, mixing of heat across zones will decay and the planet's temperature is predicted to increase.

Numerical simulations of the atmosphere are expensive due to the large number of parameters and the relatively small scales that need to be resolved. For climate models and medium-range forecast models, the governing 3D compressible Euler equations are simplified using the hydrostatic approximation (in which only the pressure gradient and the gravitational forces are retained in the vertical-momentum equation) and the anelastic approximation (in which $d\rho/dt$ is neglected), to obtain the primitive equations. Additional vertical averaging yields the shallow-water equations. One big hurdle is to accurately incorporate the effect of clouds, which is significant and is usually treated using subgrid models.

Vortices in Superfluids and Superconductors

At temperatures below 2.2 K, liquid helium is a superfluid, meaning that it acts essentially like a fluid with zero viscosity governed by the Euler equations. The fluid is irrotational, except for extremely thin vortex filaments, which are formed by quantum-mechanical processes. Since the vortices cannot end in the interior of the flow, they can be generated only at the surface or they nucleate as

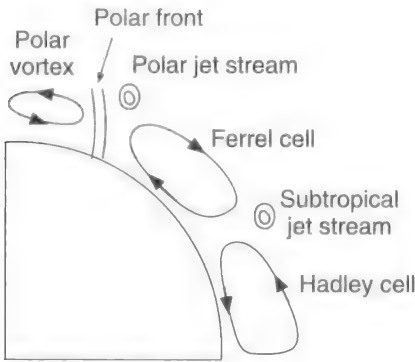


Figure 13 Vortices in the atmosphere.

vortex rings inside the fluid. As an example, if a cylindrical container with helium is rotated sufficiently fast, vortex lines attached to both ends of the container appear. These quantum vortices have discrete values of circulation ($=nh/m$, where h =Planck's constant, m =mass of helium atom, n =integer), core sizes of about 1 Å (roughly the diameter of a single hydrogen atom) and move without viscosity.

Similarly, certain types of materials lose their electric resistance at low temperatures and become superconductors. One distinguishes type-I superconductors (most pure metals) from type-II superconductors (alloys). Using the Ginzburg–Landau theory it has been predicted that in type-II superconductors a lattice of vortex filaments forms, each carrying a quantized amount of magnetic flux. This was subsequently confirmed by experimental observation. More precisely, for temperatures T below a critical value T_c , there are three regions corresponding to increasing values of the magnetic field (Figure 14). At low magnetic fields ($H < H_{c1}$), no vortices exist (superconducting phase). At intermediate values ($H_{c1} < H < H_{c2}$), the magnetic field penetrates the superconductor in the form of quantized vortices, also called flux lines (mixed phase). The values $H_{c1,c2}$ are determined by the London penetration depth λ , which measures the electromagnetic response of the superconductor. With increasing magnetic field, the density of flux lines increases until the vortex cores overlap when the upper critical field H_{c2} is reached, beyond which one recovers the normal metallic state (normal conductor).

When an external current density j is applied to the vortex system, the flux lines start to move under the action of the Lorentz force. As a result, a dissipating electric field E appears that is parallel to j , and the superconducting property of dissipation-

free current flow is lost. In order to recover the desired property of dissipation-free flow, flux lines have to be pinned, for example, by introducing inhomogeneities and structural defects. For a given pinning force, flux lines remain pinned as long as the current density stays below a critical value. A major research objective is to optimize the pinning force in order to preserve superconductivity at larger current densities.

Numerical Vortex Methods

Many numerical methods used to compute fluid flow are Eulerian schemes based on a fixed mesh, such as finite difference, finite element, and spectral methods, commonly used for example in atmosphere and ocean modeling. This section briefly describes alternative vorticity-tracking methods used to simulate incompressible inviscid vortex flows, and concludes with some extensions to viscous flows. The premise of these methods is that since the fluid velocity is determined by the vorticity through the Biot–Savart law (eqn [6]), it suffices to track only that portion of the fluid carrying nonzero vorticity. This region is often much smaller than the total fluid volume, and computational efficiency is gained. Numerical vortex methods are typically Lagrangian, that is, the computational elements move with the fluid velocity.

Point-Vortex Approximation in 2D

To compute the evolution of a vorticity distribution $\omega(\mathbf{x}, t)$ in 2D, the simplest approach is to approximate the vorticity by a set of point vortices at $\mathbf{x}_j(t)$ with circulation Γ_j and evolve them under their self-induced motion. The values Γ_j are an estimate of the initial circulation around $\mathbf{x}_j(0)$. The vortex positions $\mathbf{x}_j(t)$ evolve in the induced velocity field

$$\frac{d\mathbf{x}_j}{dt} = \sum_{\substack{k=1 \\ k \neq j}}^N \Gamma_k \mathbf{K}_{2D}(\mathbf{x}_j - \mathbf{x}_k) \tag{10}$$

where the exclusion $k \neq j$ accounts for the fact that a point vortex induces zero velocity on itself. The solution to the system of ordinary differential equations [10] can be obtained using any method, such as Runge–Kutta or Adams–Bashforth.

The point-vortex approximation can be written in Hamiltonian form as

$$\frac{dx_j}{dt} = -\frac{1}{\Gamma_j} \frac{\partial H}{\partial y_j}, \quad \frac{dy_j}{dt} = \frac{1}{\Gamma_j} \frac{\partial H}{\partial x_j} \tag{11}$$

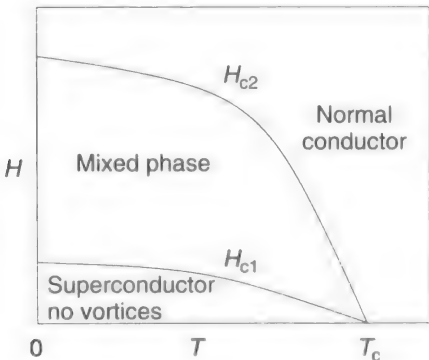


Figure 14 Superconductor phase dependence on magnetic field H and temperature T .

where the Hamiltonian

$$H(x, y) = \frac{1}{4\pi} \sum_{j=1}^N \sum_{k=1, k \neq j}^N \Gamma_j \Gamma_k \log \left[(x_j - x_k)^2 + (y_j - y_k)^2 \right] \quad [12]$$

is conserved along fluid particles, $dH/dt=0$. The method also conserves the fluid circulation and the linear and angular momenta.

Ideally, the solution to [10] should converge as $N \rightarrow \infty$ to the solution of the Euler equations. This is true for smooth vorticity distributions, but for singular distributions such as a vortex sheet, the situation is more complicated. The vortex sheet, a curve in the plane, develops a singularity in finite time at which the curvature becomes unbounded at a point. The point-vortex approximation converges before the singularity formation time, provided the growth of spurious roundoff error due to Kelvin–Helmholtz instability is suppressed using a filter. However, past the singularity formation time, the point-vortex approximation no longer converges.

The general approach is to replace the singular kernel K_{2D} by a regularization K_{2D}^δ , such as

$$K_{2D}^\delta = \frac{1}{2\pi} \frac{-y\mathbf{i} + x\mathbf{j}}{|\mathbf{x}|^2 + \delta^2} \quad [13a]$$

$$K_{2D}^\delta = \frac{1}{2\pi} \frac{-y\mathbf{i} + x\mathbf{j}}{|\mathbf{x}|^2} \left(1 - e^{-|\mathbf{x}|^2/\delta^2} \right) \quad [13b]$$

where δ is a numerical parameter. The regularization amounts to replacing the δ -function vorticity of a point vortex by an approximate δ -function. In order to recover the solution to the Euler equations, it is necessary to study the limit $N \rightarrow \infty, \delta \rightarrow 0$. For smooth vorticity distributions, this process converges. For vortex sheet initial data, there is evidence of convergence, but details of the limiting behavior remain under investigation. Regularized solutions with fixed value δ and vortex sheet initial data are shown in Figures 6 and 15. Figure 6 shows the onset of the Kelvin–Helmholtz instability in a periodically perturbed flat vortex sheet. Figure 15 shows the rollup of an elliptically loaded flat vortex sheet that models the evolution of an aircraft wake (see Figure 7). The correspondence between the two-dimensional simulation and the three-dimensional wake is made by replacing the spatial coordinate in the aircraft's line of flight by a time coordinate.

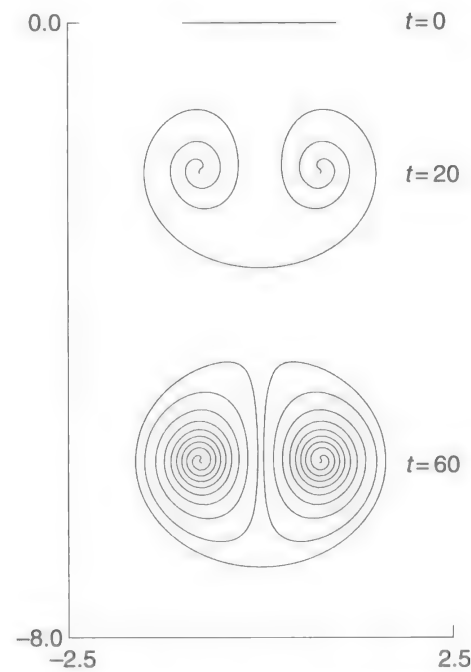


Figure 15 Computed evolution of an elliptically loaded flat vortex sheet.

Contour Dynamics in 2D

Consider a planar patch of constant vorticity ω_0 bounded by a curve $\mathbf{x}(s, t), 0 \leq s \leq L$, moving in inviscid, incompressible flow. In view of Kelvin's theorem and eqn [5], the vorticity in the patch remains constant and equal to ω_0 for all time, and the patch area remains constant. Only the patch boundary moves. The velocity at a point $\mathbf{x}(\alpha, t)$ on the boundary can be written as a line integral over the boundary:

$$\frac{d\mathbf{x}}{dt} = -\frac{\omega_0}{2\pi} \int_C \log |\mathbf{x} - \mathbf{x}(s, t)| \frac{\partial \mathbf{x}}{\partial s} ds \quad [14]$$

The contour dynamics method consists of approximating a given vorticity distribution by a superposition of vortex patches, and moving their boundaries according to eqn [14]. This method has been applied to compute the evolution of single-vortex patches and shear layers, and to geophysical flows. Typically, filamentation occurs: the patch develops thin filaments which increase the boundary length significantly and thereby the computational expense. The approach generally taken is to remove the thin filaments at several times throughout the computation, which is referred to as contour surgery. The contour dynamics approach as well as the point-vortex approximation have also been generalized to treat quasigeostrophic flows.

Vortex Filament Methods in 3D

Vortex simulations in 3D differ from those in 2D in that the stretching term in eqn [4] needs to be incorporated. The vortex filament method approximates the fluid vorticity by a finite number of filaments whose circulation remains constant in time. Each filament is marked by computational mesh points which move with the regularized induced velocity. The regularization is necessary to prevent the infinite self-induced velocities of curved vortex filaments. As in 2D, this method automatically conserves circulation. Vorticity stretching is accounted for by the stretching between computational mesh points. As the filament length increases, more meshpoints are typically introduced to keep it resolved. Also, the number of filaments can be increased throughout the simulation to maintain resolution.

Viscous Vortex Methods

While inviscid models are expected to approximate small viscosity fluids well far from boundaries, near boundaries, where vortex shedding is an inherently viscous mechanism, it is important to incorporate the effects of viscosity. The first methods to do so used operator splitting in which inviscid and viscous terms of the Navier–Stokes equations were solved in a sequential manner. In each time step, the computational elements would first be convected, and then they would be diffused by a random-walk scheme. The particle strength exchange method, introduced more recently, does not rely on operator splitting and has better accuracy. The particle position and vorticity evolve simultaneously, and viscous diffusion is accounted for in a consistent manner.

Vortex dynamics continues to be a source of interesting problems of theoretical and practical importance. In particular, much remains to be learned to better understand turbulence and the transition to turbulence, a process dominated by deterministic vortex dynamics.

Further Remarks

Finally, some remarks on relevant literature on this subject are in order. Lugt (1983) and Tritton (1988) are recommended as elementary introduction to vortex flows. van Dyke (1982) presents beautiful and instructive flow visualizations. Comprehensive treatments of incompressible fluid dynamics are given in Batchelor (1967), Chorin and Marsden (1992), Lamb (1932), and Saffman (1992), and compressible flow is treated in Anderson (1990). Cottet and Koumoutsakos (2000) give an overview of numerical vortex methods.

Special topics have also been addressed; atmosphere (Andrews *et al.* 1987), point vortex motion and chaos (Aref 1983, Newton 2001, Ottino 1989), superfluids and superconductors (Blatter *et al.* 1994, Donnelly 1991), turbulence theory using statistical mechanics (Chorin 1994), vortex reconnection (Kida and Takaoka 1994), theory for Euler and Navier–Stokes equations (Majda and Bertozzi 2002), contour dynamics (Pullin 1992), vortex rings (Shariff and Leonard 1992), and aircraft trailing vortices (Spalart 1998). Green (1995) includes survey articles on various topics.

Nomenclature

a	vortex ring core size
\mathbf{g}	gravitational field
H	Hamiltonian
K_{2D}	singular velocity kernel
$K_{2D,\delta}$	regularized velocity kernel
$\rho(\mathbf{x}, t)$	fluid density
R	vortex ring radius
$T(\mathbf{x}, t)$	temperature
\mathbf{U}	translation velocity
$\mathbf{u}(\mathbf{x}, t) = u(\mathbf{x}, t)\mathbf{i} + v(\mathbf{x}, t)\mathbf{j} + w(\mathbf{x}, t)\mathbf{k}$	fluid velocity
$w(k)$	dispersion relation
$\boldsymbol{\omega} = \nabla \times \mathbf{u}$	vorticity
$\omega = v_x - u_y$	scalar vorticity
Γ	ring circulation

See also: Abelian Higgs Vortices; Incompressible Euler Equations: Mathematical Theory; Integrable Systems: Overview; Interfaces and Multicomponent Fluids; Intermittency in Turbulence; Newtonian Fluids and Thermohydraulics; Point-Vortex Dynamics; Stochastic Hydrodynamics; Superfluids; Topological Knot Theory and Macroscopic Physics; Turbulence Theories.

Further Reading

- Anderson JD (1990) *Modern Compressible Flow with Historical Perspective*, 2nd edn. New York: McGraw-Hill.
- Andrews DG, Holton JR, and Leovy CB (1987) *Middle Atmosphere Dynamics*. Orlando: Academic Press.
- Aref H (1983) Integrable, chaotic, and turbulent vortex motion in two-dimensional flows. *Annual Review of Fluid Mechanics* 15: 345–389.
- Batchelor GK (1967) *An Introduction to Fluid Dynamics*. Cambridge: Cambridge University Press.
- Blatter G, Feigel'man MV, Geshkenbein VB, Larkin AI, and Vinokur VM (1994) Vortices in high-temperature superconductors. *Reviews of Modern Physics* 66(4): 1125–1388.
- Chorin AJ (1994) *Vorticity and Turbulence*. New York: Springer.
- Chorin AJ and Marsden JE (1992) *A Mathematical Introduction to Fluid Mechanics*, 3rd edn. New York: Springer.
- Cottet G-H and Koumoutsakos PD (2000) *Vortex Methods: Theory and Practice*. Cambridge: Cambridge University Press.

- Donnelly RJ (1991) *Quantized Vortices in Helium II*. Cambridge: Cambridge University Press.
- Green SI (ed.) (1995) *Fluid Vortices*. Dordrecht: Kluwer Academic.
- Kida S and Takaoka M (1994) Vortex reconnection *Annual Review of Fluid Mechanics* 26: 169–189.
- Lamb H (1932) *Hydrodynamics*, 6th edn. New York: Dover.
- Lugt HJ (1983) *Vortex Flow in Nature and Technology*. New York: Wiley.
- Majda AJ and Bertozzi AL (2002) *Vorticity and Incompressible Flow*. Cambridge: Cambridge University Press.
- Newton PK (2001) *The N-Vortex Problem: Analytical Techniques*. New York: Springer.
- Pullin DI (1992) Contour dynamics methods. *Annual Review of Fluid Mechanics* 24: 89–115.
- Ottino JM (1989) *The Kinematics of Mixing: Stretching, Chaos, and Transport*. Cambridge: Cambridge University Press.
- Saffman PG (1992) *Vortex Dynamics*. Cambridge: Cambridge University Press.
- Shariff K and Leonard A (1992) Vortex rings. *Annual Review of Fluid Mechanics* 24: 235–279.
- Spalart PR (1998) Airplane trailing vortices. *Annual Review of Fluid Mechanics* 30: 107–138.
- Tritton DJ (1988) *Physical Fluid Dynamics*, 2nd edn. Oxford: Clarendon Press.
- van Dyke M (1982) *Album of Fluid Motion*. Stanford: The Parabolic Press.

Vortices see Abelian Higgs Vortices; Point-Vortex Dynamics

W

Wave Equations and Diffraction

M E Taylor, University of North Carolina, Chapel Hill, NC, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The most basic wave equation is

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = 0 \quad [1]$$

for $u = u(t, x)$, where Δ is the Laplace operator, given by $\Delta u = \partial^2 u / \partial x_1^2 + \dots + \partial^2 u / \partial x_n^2$ on n -dimensional Euclidean space \mathbb{R}^n . More generally, u might be defined on $\mathbb{R} \times M$, where \mathbb{R} is the t -axis and M is a Riemannian manifold, with a metric tensor given in local coordinates by (g_{jk}) . Then the Laplace–Beltrami operator is given, in local coordinates, by

$$\Delta u = g^{-1/2} \sum_{j,k} \frac{\partial}{\partial x_j} \left(g^{1/2} g^{jk} \frac{\partial u}{\partial x_k} \right) \quad [2]$$

where (g^{jk}) is the matrix inverse to (g_{jk}) and $g = \det(g_{jk})$. Even if one concentrates on wave propagation in Euclidean space, one frequently wants to use curvilinear coordinates, and [2] is useful. Equation [1] is supplemented by initial conditions of the form

$$u(0, x) = f(x), \quad \partial_t u(0, x) = g(x) \quad [3]$$

called Cauchy data. If the spatial domain M has a boundary ∂M (e.g., if M is a bounded region in \mathbb{R}^n), then boundary conditions are imposed. The most common are the Dirichlet boundary condition

$$u(t, x) = 0 \quad \text{for } x \in \partial M \quad [4]$$

and the Neumann boundary condition

$$\partial_\nu u(t, x) = 0 \quad \text{for } x \in \partial M \quad [5]$$

where $\partial_\nu u$ denotes the normal derivative of u at the boundary. More generally, one might have a driving force, and replace 0 on the right-hand side of [1] by a function $F(t, x)$. Similarly, one can consider nonzero boundary data in [4] and [5].

The wave equation [1] models a number of physical phenomena, at least in the linear approximation. The vibration of a drum head is modeled by [1], with M a planar domain, and with the Dirichlet boundary condition [4]. The motion of sound waves in a room with hard walls is modeled by [1], with M a region in \mathbb{R}^3 , and with the Neumann boundary condition [5]. The propagation of electromagnetic waves is given by Maxwell's equations:

$$\begin{aligned} \frac{\partial E}{\partial t} - \text{curl } B &= -J \\ \frac{\partial B}{\partial t} + \text{curl } E &= 0 \\ \text{div } E &= \rho \\ \text{div } B &= 0 \end{aligned} \quad [6]$$

where ρ is the electric charge density and J the current. These equations yield [1] (with the right-hand side replaced by some function $F(t, x)$ if J and ρ are not zero) for the components of the electric field E and the magnetic field B . If the propagation is in a region M in \mathbb{R}^3 bounded by a perfect conductor, then the boundary conditions are that E is normal to ∂M and B is tangential to ∂M . If ∂M is flat, these equations can be decomposed into Dirichlet problems for some components and Neumann problems for the rest, but if ∂M is curved such a decomposition is not possible.

Other models of vibrating objects produce variants of [1]. Examples include vibrating elastic solids, yielding an equation like [1] with Δu replaced by $\mu \Delta u + (\lambda + \mu) \text{grad div } u$, for linear elasticity. Here Δ acts componentwise on u , and μ and λ are constants, called Lamé constants. Other examples model vibrations of crystals and propagation of electromagnetic waves in crystals. Further interesting phenomena arise in these various cases, such as Rayleigh waves in linear elasticity and conical refraction in crystal optics.

Here we discuss the propagation of waves and their reflection and diffraction at boundaries. In the interest of providing reasonable coverage in a brief space, we restrict attention to the wave equation [1].

Basic Propagation Phenomena

The simplest examples of waves propagating according to [1] are plane waves, of the form

$$u(t, x) = \varphi(x \cdot \omega - t) \tag{7}$$

with $(t, x) \in \mathbb{R} \times \mathbb{R}^n$, ω a unit vector in \mathbb{R}^n , and φ a function on \mathbb{R} . If φ has two continuous derivatives, [7] defines a classical solution of [1]. More generally, one can allow φ to be less regular. For example, it could be piecewise smooth with a jump discontinuity at some point $a \in \mathbb{R}$. In such a case, u will be piecewise smooth with a jump across the n -dimensional surface $x \cdot \omega - t = a$ in $\mathbb{R} \times \mathbb{R}^n$, which will solve [1] in a weak, or distributional, sense. For fixed t , $u(t, \cdot)$ has a jump across the $(n - 1)$ -dimensional surface $\Sigma_t = \{x \in \mathbb{R}^n : x \cdot \omega = t + a\}$. As t varies, Σ_t moves in the direction ω with unit speed.

There are also spherical wave solutions to [1] on $\mathbb{R} \times \mathbb{R}^n$, such as

$$u(t, x) = \frac{\text{sgn } t}{2\pi} (t^2 - |x|^2)_+^{-1/2} \tag{8}$$

for $n = 2$, and

$$u(t, x) = \frac{1}{4\pi t} \delta(|x| - |t|) \tag{9}$$

for $n = 3$. Here $s_+ = s$ for $s > 0$, $s_+ = 0$ for $s \leq 0$, and $\delta(s)$ is the Dirac delta function. In fact, [8] and [9] are “fundamental solutions” (more on which in the section on harmonic analysis) to the wave equation on $\mathbb{R} \times \mathbb{R}^n$, for $n = 2$ and 3 , respectively. In such cases, the singularity in $u(t, \cdot)$ for each fixed t lies in $\Sigma_t = \{x \in \mathbb{R}^n : |x| = |t|\}$, a family of surfaces in \mathbb{R}^n that moves, in the direction of the normal to Σ_t , at unit speed.

The examples mentioned above illustrate two general phenomena about the behavior of solutions to [1]. The first is finite propagation speed. Its general formulation is that, given a closed set $K \subset M$,

$$\begin{aligned} \text{supp } f, g \subset K &\Rightarrow \text{supp } u(t, \cdot) \\ &\subset \{x \in M : \text{dist}(x, K) \leq |t|\} \end{aligned} \tag{10}$$

In fact, given that [8]–[9] are fundamental solutions, [10] is a consequence of these formulas when $M = \mathbb{R}^2$ or \mathbb{R}^3 . The result [10] is true in great generality, with well-known demonstrations involving energy estimates.

The second phenomenon involves propagation of singularities. Typically, if the Cauchy data f and g in [3] are smooth on the complement of an $(n - 1)$ -dimensional surface Σ_0 , perhaps with a jump across Σ_0 , or such a singularity as in [8] or [9], the solution $u(t, x)$ will be a sum of two terms, with singularities

of a similar nature on the surfaces Σ_t^\pm , moving at unit speed in the direction of their normals, Σ_t^+ flowing from Σ_0 in one direction and Σ_t^- in the other. This also holds for the manifold case [2]. That happens at least until such surfaces develop singularities, when matters become more elaborate.

An alternative way to describe how the set of singularities evolves is the following. Let S_1M denote the space of unit vectors tangent to M ; this is a submanifold of the tangent bundle of M , TM . There is a natural projection $\pi : S_1M \rightarrow M$. Associated to a smooth surface Σ_0 of dimension $n - 1$ in M (of dimension n) are two preimages Λ_0^+ and Λ_0^- in S_1M , consisting of unit vectors lying over points of Σ_0 and normal to Σ_0 . The geodesic flow is a flow on S_1M , and it takes Λ_0^\pm to smooth $(n - 1)$ -dimensional surfaces Λ_t^\pm in S_1M . The sets Σ_t^\pm are the images of Λ_t^\pm under π . The geodesics starting out at points in Λ_0^\pm and sweeping out Λ_t^\pm are the rays along which the singularities of the solution u propagate.

This latter description works for all t if M has no boundary and is complete, that is, all geodesics are defined for all t , although singularities develop in the images $\pi(\Lambda_t^\pm) = \Sigma_t^\pm$, at points $p \in \Sigma_t^\pm$, where Λ_t^\pm meets T_pM nontransversally. The behavior of u near such singular points of Σ_t^\pm , known as caustics, is more complicated than that near regular points, but it can be captured in terms of integrals. Methods of establishing this propagation of singularities are discussed in the section on geometrical optics.

Such a description needs further elaboration if M has a boundary. One of the principal problems of diffraction theory is to explain how singularities of solutions to [1], with a boundary condition such as [4] or [5], propagate and reflect off the boundary.

Considering the case where M is a half-space in \mathbb{R}^n ,

$$M = \mathbb{R}_+^n = \{x \in \mathbb{R}^n : x_n \geq 0\} \tag{11}$$

provides a guide to the simplest reflection phenomena. In such a case, one can solve the Dirichlet or Neumann boundary problem for the wave equation [1] by the method of images. One extends f and g from \mathbb{R}_+^n to \mathbb{R}^n . For the Dirichlet problem [4], one takes odd extensions, $f(x', -x_n) = -f(x', x_n)$, and similarly for g . For the Neumann problem [5], one takes even extensions, $f(x', -x_n) = f(x', x_n)$, etc. One then solves the wave equation [1] on $\mathbb{R} \times \mathbb{R}^n$ with the extended Cauchy data, and the restriction to $\mathbb{R} \times \mathbb{R}_+^n$ solves the respective boundary problem. Suppose Σ_0 is a smooth $(n - 1)$ -dimensional surface that does not meet $\partial\mathbb{R}_+^n$, and that f and g have singularities on Σ_0 , as above. (Suppose for simplicity that f and g vanish near $\partial\mathbb{R}_+^n$.) Those rays issuing

from normals to Σ_0 have mirror images, which are rays in \mathbb{R}^n_- . If such a ray hits $\partial\mathbb{R}^n_+$, its mirror image does so also, and continues into \mathbb{R}^n_+ , as the reflected ray. The singularities of u propagate along such reflected rays.

Such a description extends to a general complete Riemannian manifold with boundary M , in the case of rays that hit the boundary transversally. Such a ray is reflected by retaining the tangential component of its velocity vector at the point of intersection ∂M and reversing the sign of the normal component. One says that the ray is reflected according to the laws of geometrical optics. Singularities of u carried by such rays that hit ∂M are correspondingly reflected. Methods to establish such transversal reflection of singularities are natural extensions of those developed to treat the propagation away from ∂M , mentioned above.

Matters become more delicate when there are rays that are tangent to ∂M . A model example is given by

$$M = \mathbb{R}^n \setminus B, \quad B = \{x \in \mathbb{R}^n : |x| < 1\} \quad [12]$$

which one takes when studying the scattering of waves in \mathbb{R}^n by the obstacle B . Consider a solution to [1] with boundary condition given by [4] or [5] that has a simple singularity on $\Sigma_t = \{x \in \mathbb{R}^n : x_n = t\}$ for $t < -1$. The associated rays are of the form $\gamma_{x'}(t) = (x', t)$, for $t < -1$, with $x' \in \mathbb{R}^{n-1}$. If $|x'| > 1$, these rays continue on in $\mathbb{R}^n \setminus B$, for all $t \geq -1$. If $|x'| < 1$, these rays hit $\partial M = \partial B$ transversally, and their reflection is as described above. If $|x'| = 1$, these rays hit ∂B tangentially, at $t = 0$; they are sometimes called grazing rays. One also continues them past $t = 0$. One defines in this fashion Σ_t for $t \geq -1$. The region

$$\mathcal{S} = \{x = (x', x_n) \in \mathbb{R}^n \setminus B : |x'| < 1, x_n > 0\} \quad [13]$$

is called the “shadow region.” It is disjoint from Σ_t for all t . The solution u is smooth in \mathcal{S} for all t , although it is not identically zero. The set

$$\mathcal{S}^b = \{x = (x', x_n) \in \mathbb{R}^n \setminus B : |x'| = 1, x_n \geq 0\} \quad [14]$$

is the “shadow boundary.”

One can replace B in [12] by a more general smooth, convex obstacle K , with positive Gauss curvature everywhere, and the same considerations of transversal and grazing rays and shadow regions apply. These notions also extend to a more general class of Riemannian manifolds with boundary, called manifolds with diffractive boundary. In the case $K = B$, one can use separation of variables to reduce the problem of analyzing solutions to [1] and showing that singularities propagate along such rays to a problem in harmonic analysis on the sphere

S^{n-1} . For more general convex obstacles K or manifolds with diffractive boundary, other techniques are required, to show that waves reflect off the boundary in a fashion similar to the case [12].

Another situation arises if instead of [12] one takes $M = B$, or more generally $M = K$, a convex region as described above. A ray starting off from a point in ∂M , almost tangent to ∂M but with a small component in the direction of the normal pointing into M , will undergo many reflections in a short time. Upon shrinking the normal component of the initial velocity to zero, one obtains in the limit a geodesic in ∂M , known as a gliding ray. In such a case, singularities of solutions to [1], with such a boundary condition as [4] or [5], propagate along both transversally reflected and gliding rays.

For the generic smooth obstacle K in \mathbb{R}^n , the second fundamental form can have a variety of signatures at various boundary points. Various types of “generalized rays” occur – generally speaking limits of sequences of transversally reflected rays. This situation also holds for general complete Riemannian manifolds with smooth boundary. The main result about propagation of singularities in such a case is that it is always along such generalized rays. This was established by Melrose and Sjöstrand (1978).

Further diffraction effects arise when ∂M has singularities, such as edges and corners. The simplest example is

$$M = \{x \in \mathbb{R}^2 : a \leq \theta \leq b, r \geq 0\} \quad [15]$$

where (r, θ) are the polar coordinates of $x \in \mathbb{R}^2$, and we assume $0 \leq a < b \leq 2\pi$. Here one is studying the diffraction of waves by a wedge. In the limiting case $a = 0$, $b = 2\pi$, the wedge becomes a half-line, that is,

$$M = \mathbb{R}^2 \setminus \{(x_1, 0) : x_1 > 0\} \quad [16]$$

Singularities of solutions to [1] on $\mathbb{R} \times M$ with such a boundary condition as [4] or [5] propagate in the interior of M and reflect off the regular points of ∂M as before. If a family of continuous, piecewise smooth curves Σ_t carrying the singularity of u hit the corner $x = 0$ at $t = a$, this reflection creates a tear in Σ_t for $t > a$. In addition, a diffracted wave spreads out from the corner at unit speed. This diffracted wave carries a singularity that is weaker than that of the incident wave. For example, if one has a solution like [8], but shifted to have support in a disk of radius $|t|$ about a point $p \neq 0$ in \mathbb{R}^2 , for small $|t|$, then the diffracted wave will have a jump discontinuity.

The space M in [15] is a special case of a cone. More generally, if N is a complete Riemannian

manifold (possibly with boundary), then the cone $C(N)$ with base N is the set

$$C(N) = [0, \infty) \times N \tag{17}$$

with all points $(0, x), x \in N$, identified, with the metric tensor

$$ds^2 = dr^2 + r^2g \tag{18}$$

where g is the metric tensor on N , and points on $C(N)$ are denoted $(r, x), r \in [0, \infty), x \in N$. The space in [14] has the form $M = C(N)$ with $N = [a, b]$, an interval. A cone in Euclidean space \mathbb{R}^n is of the form $C(N)$ with N a domain in the unit sphere S^{n-1} .

The propagation of singularities for solutions to [1] on $C(N)$, when N has smooth boundary, has a description similar to that above for the case [15]. Again, there is a diffracted wave set off from the conic point $\{r=0\}$ when a singularity of a wave hits it. The diffracted wave is typically $(n-1)/2$ units smoother than the singular wave producing it, where $n = \dim C(N)$. For example, the fundamental solution to the wave equation on $C(N)$ produces a diffracted wave which is the sum of a jump discontinuity and (in general) a logarithmic singularity.

In fact, precise understanding of the behavior of the fundamental solution to the wave equation on $C(N)$ is encoded in terms of the behavior of the solution operator to the wave equation on the base N . This is discussed in further detail in the section on harmonic analysis. In the case where $C(N)$ is given by [15], we are dealing with the wave equation on an interval $[a, b]$, whose behavior is elementary.

One can use analysis of [15] together with finite propagation speed to get a good qualitative picture of diffraction of waves in \mathbb{R}^2 by a polygonal obstacle. A variation of this argument allows one to understand the behavior of the wave equation on a “polygonal” domain N in S^2 , that is, one whose boundary consists of a finite number of geodesic segments in S^2 . Going from there to $C(N)$, one can then analyze diffraction of waves in \mathbb{R}^3 by a polyhedron.

It is worth remarking how the “shadow region” for such an obstacle as a wedge in \mathbb{R}^2 differs from that in [12]–[14]. For example, if one considers M given by [16] and $u(t, x) = \delta(x_2 - t)$, for $t < 0$, then the region

$$S = \{x = (x_1, x_2) : x_1, x_2 > 0\} \tag{19}$$

is the “shadow region,” in the sense that rays either missing or reflecting off the obstacle $\{(x_1, 0) : x_1 > 0\}$ do not enter the region [19]. However, unlike the case [13], the solution $u(t, x)$ is not smooth in the region [19] for $t > 0$. There is a singularity there,

although it is weaker than the singularity of the main wave.

Taking Cartesian products of spaces of the form [15] with \mathbb{R}^k yields spaces with k -dimensional edges. There are also spaces with curvy edges. Rather than continuing with further general description, one more particular case is discussed next, which has had a historical significance. Namely, we consider the reflection of waves in \mathbb{R}^3 off a disk, that is, take

$$M = \mathbb{R}^3 \setminus D, \quad D = \{(x_1, x_2, 0) : x_1^2 + x_2^2 \leq 1\} \tag{20}$$

Consider a wave given for $t < 0$ by $u(t, x) = \delta(x_3 - t)$. This wave hits $D = \partial M$ at $t = 0$, giving off a diffracted wave, traveling away from $\sigma = \{(x_1, x_2, 0) : x_1^2 + x_2^2 = 1\}$ at speed 1 for $t > 0$. This diffracted wave carries a singularity that blows up like the $-1/2$ power of the distance to the torus of points of distance t from σ , for $t \in (0, 1)$. For $t > 1$, there is a focusing effect along the x_3 -axis, producing a stronger singularity for $u(t, x)$ there.

This sort of phenomenon was understood, at least from a heuristic point of view, in the nineteenth century, and it played a role in an important argument of Poisson. At the time, there was a debate about whether the propagation of light was a wave phenomenon. Poisson did not think it was, and he noted that if it were, the light waves propagated past such an obstacle should produce a bright spot along the axis normal to the disk and through its center. The experiment was performed and the bright spot was observed. This is now called the Poisson spot, and its occurrence convinced many physicists, including Poisson, that the propagation of light is a wave phenomenon.

Harmonic Analysis and the Wave Equation

The wave equation [1] with Cauchy data [3] can be regarded as an operator differential equation, with solution

$$u(t, x) = \cos t\sqrt{-\Delta} f(x) + \frac{\sin t\sqrt{-\Delta}}{\sqrt{-\Delta}} g(x) \tag{21}$$

This brings one to investigate functions of the self-adjoint operator Δ . If $M = \mathbb{R}^n$, one can do this using the Fourier transform, which is given by

$$\mathcal{F}f(\xi) = \hat{f}(\xi) = (2\pi)^{-n/2} \int f(x)e^{-ix \cdot \xi} dx \tag{22}$$

One defines \mathcal{F}^* by changing $e^{-ix \cdot \xi}$ to $e^{ix \cdot \xi}$ in [22], and the Fourier inversion formula says \mathcal{F} and \mathcal{F}^* are

inverses of each other on various function spaces, including $L^2(\mathbb{R}^n)$. Then one has

$$\varphi(\sqrt{-\Delta})f(x) = (2\pi)^{-n/2} \int \varphi(|\xi|)\hat{f}(\xi)e^{ix\cdot\xi} d\xi \quad [23]$$

Note that [23] is equal to

$$\int \Phi(x-y)f(y) dy = \Phi * f(x) \quad [24]$$

where

$$\Phi(x) = (2\pi)^{-n} \int \varphi(|\xi|)e^{ix\cdot\xi} d\xi \quad [25]$$

In particular, [21] becomes

$$u(t, x) = \frac{\partial}{\partial t} R_t * f(x) + R_t * g(x) \quad [26]$$

where

$$R_t(x) = (2\pi)^{-n} \int \frac{\sin t|\xi|}{|\xi|} e^{ix\cdot\xi} d\xi \quad [27]$$

is the fundamental solution to the wave equation.

The integral [27] is not an easy integral when $n > 1$, but the answer can be derived by analytic continuation from the Poisson kernel, that is,

$$\begin{aligned} e^{-y\sqrt{-\Delta}}f(x) &= P_y * f(x) \\ P_y(x) &= C_n y(|x|^2 + y^2)^{-(n+1)/2} \end{aligned} \quad [28]$$

where $C_n = \pi^{-(n+1)/2} \Gamma((n+1)/2)$. One gets

$$R_t(x) = \lim_{\varepsilon \searrow 0} \frac{C_n}{n-1} \operatorname{Im}(|x|^2 - (t-i\varepsilon)^2)^{-(n-1)/2} \quad [29]$$

Taking this limit for $n=2, 3$ yields the formulas [8]–[9]. There are several ways to derive [28]. One, which is flexible and useful for other situations, derives it from the formula for the heat kernel,

$$e^{t\Delta}f(x) = H_t * f(x), \quad H_t(x) = (4\pi t)^{-n/2} e^{-|x|^2/4t} \quad [30]$$

via the subordination identity:

$$\begin{aligned} e^{-yA} &= \frac{y}{2\pi^{1/2}} \int_0^\infty e^{-y^2/4t} e^{-tA^2} t^{-3/2} dt \\ A &> 0, \quad y > 0 \end{aligned} \quad [31]$$

with $A = \sqrt{-\Delta}$. The heat kernel can be computed via [23], which becomes a well-known Gaussian integral. The identity [31] can be proved using the fact that the Fourier integral formula for $P_y(x)$ is elementary to compute when $n=1$.

To understand functions of the Laplace operator on a cone $C(N)$, one uses

$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{n-1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \Delta_N \quad [32]$$

where Δ_N is the Laplace operator on N , which follows from [2] and [18]. Here $n = \dim C(N)$. This is a modified Bessel operator. We define the operator

$$\nu = (-\Delta_N + \alpha^2)^{1/2}, \quad \alpha = -\frac{n-2}{2} \quad [33]$$

For each ν_j in the spectrum of ν , we consider the Hankel transform

$$H_{\nu_j} g(\lambda) = \int_0^\infty g(r) J_{\nu_j}(\lambda r) r dr \quad [34]$$

where J_{ν_j} is the Bessel function of order ν_j . The Hankel inversion formula says H_{ν_j} is unitary on $L^2(\mathbb{R}^+, r dr)$, and is its own inverse. Consequently, we can write the action of $\varphi(\sqrt{-\Delta})$ on $L^2(C(N))$ as

$$\varphi(\sqrt{-\Delta})g(r, x) = \int_0^\infty K_\varphi(r, s, \nu) g(s, x) s^{n-1} ds \quad [35]$$

where $K_\varphi(r, s, \nu)$ is a family of operators on $L^2(N)$, given by

$$K_\varphi(r, s, \nu) = (rs)^\alpha \int_0^\infty \varphi(\lambda) J_\nu(\lambda r) J_\nu(\lambda s) \lambda d\lambda \quad [36]$$

To obtain the wave kernel on $C(N)$, one can analytically continue formulas for the Poisson kernel, for $e^{-y\sqrt{-\Delta}}$. Such formulas arise from the Lipschitz–Hankel identity:

$$\begin{aligned} &\int_0^\infty e^{-y\lambda} J_\nu(r\lambda) J_\nu(s\lambda) d\lambda \\ &= \frac{1}{\pi} (rs)^{-1/2} Q_{\nu-1/2} \left(\frac{r^2 + s^2 + y^2}{2rs} \right) \end{aligned} \quad [37]$$

Here $Q_{\nu-1/2}(\tau)$ is a Legendre function. The identity [37] is one of the more difficult identities in the theory of Bessel functions. It is useful to know that it can be derived by applying a slight variant of the subordination identity [31] to the more elementary identity

$$\int_0^\infty e^{-t\lambda^2} J_\nu(r\lambda) J_\nu(s\lambda) \lambda d\lambda = \frac{1}{2t} e^{-(r^2+s^2)/4t} I_\nu \left(\frac{rs}{2t} \right) \quad [38]$$

(where $I_\nu(y) = e^{-\pi i \nu/2} J_\nu(iy)$ for $y > 0$), which describes the behavior of the heat kernel on $C(N)$.

Carrying out the analytic continuation of [37] to imaginary y yields results stated in the section on basic propagation phenomena, once one

understands the behavior of families of functions of the operator ν so produced. An approach taken by Cheeger and Taylor (1982) to this was to synthesize these operators from $e^{is\nu}$, $s \in \mathbb{R}$, and deduce their behavior from the behavior of the solution operator to the wave equation on the base N .

One can apply similar considerations to $M = \mathbb{R}^n \setminus B$, which is the truncated cone $[1, \infty) \times S^{n-1}$, with metric tensor [18], where g is the metric tensor on S^{n-1} , and Laplace operator given by [32], with Δ_N the Laplace operator on S^{n-1} . The problem of diffraction of waves by the ball B can be recast as solving

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} - \Delta u &= 0 \quad \text{on } \mathbb{R} \times M \\ u|_{\mathbb{R} \times \partial M} &= f, \quad u(t, x) = 0 \quad \text{for } t \ll 0 \end{aligned} \tag{39}$$

with f compactly supported on $\mathbb{R} \times \partial M$. Taking the partial Fourier transform with respect to t yields the reduced wave equation

$$(\Delta + \lambda^2)v = 0 \quad \text{for } |x| > 1, \quad v|_{S^{n-1}} = g(x, \lambda) \tag{40}$$

and the condition $u(t, x) = 0$ for $t \ll 0$ yields for v the outgoing radiation condition

$$r^{(n-1)/2} \left(\frac{\partial v}{\partial r} - i\lambda v \right) \rightarrow 0 \quad \text{as } r \rightarrow \infty \tag{41}$$

The solution is

$$v(x, \lambda) = r^{-(n-2)/2} \frac{H_{\nu_j}^{(1)}(\lambda r)}{H_{\nu_j}^{(1)}(\lambda)} g(x, \lambda) \tag{42}$$

with ν as in [33] and $H^{(1)}$ the Hankel function.

The behavior of $H_{\nu_j}^{(1)}(\lambda r)/H_{\nu_j}^{(1)}(\lambda)$ as $\nu_j, \lambda \rightarrow \infty$ with ratio in a small neighborhood of 1 can be shown to control the behavior of the solution u to [39] near grazing rays. There is an asymptotic formula for this, which is one of the most delicate analytical results in the theory of Bessel functions. The result is that, uniformly for z near 1, as $\mu \rightarrow \infty$,

$$\begin{aligned} H_{\mu}^{(1)}(\mu z) &\sim 2e^{-\pi i/3} \left(\frac{4\zeta}{1-z^2} \right)^{1/4} \\ &\times \left\{ A_+(\mu^{2/3}\zeta) \sum_{k \geq 0} a_k(\zeta) \mu^{1/3-2k} \right. \\ &\quad \left. + A'_+(\mu^{2/3}\zeta) \sum_{k \geq 0} b_k(\zeta) \mu^{5/3-2k} \right\} \end{aligned} \tag{43}$$

Here

$$A_+(\xi) = \text{Ai}(e^{-2\pi i/3}\xi) \tag{44}$$

where Ai is the Airy function. The coefficients $a_k(\zeta)$ and $b_k(\zeta)$ are smooth functions of their argument, $\zeta = \zeta(z)$, which is defined by

$$\frac{2}{3}\zeta^{3/2} = \int_z^1 \sqrt{1-t^2} \frac{dt}{t} \tag{45}$$

Making use of [43] in [44], one can obtain a parametrix for u (i.e., a solution modulo a C^∞ error) whose form is a special case of the formula [50], which we will present in the next section.

Geometrical Optics and Extensions

By results of the last section, the solution to [1] when $M = \mathbb{R}^n$ has the form

$$u(t, x) = \sum_{\pm} \int e^{\pm i t |\xi| + i x \cdot \xi} \hat{h}_{\pm}(\xi) d\xi \tag{46}$$

where the functions h_{\pm} are produced from the initial data via simpler transformations. For a general metric tensor, one can produce a parametrix (i.e., an approximation to $u(t, x)$ with a C^∞ error) in the following form:

$$u(t, x) = \sum_{\pm} \int a^{\pm}(t, x, \xi) e^{i\varphi^{\pm}(t, x, \xi)} \hat{h}_{\pm}(\xi) d\xi \tag{47}$$

Here the phase functions $\varphi^{\pm}(t, x, \xi)$ are smooth for $\xi \neq 0$ and homogeneous of degree 1 in ξ . The amplitudes $a^{\pm}(t, x, \xi)$ are smooth and have asymptotic expansions as $|\xi| \rightarrow \infty$:

$$a^{\pm}(t, x, \xi) \sim \sum_{k \geq 0} a_k^{\pm}(t, x, \xi) \tag{48}$$

with $a_k^{\pm}(t, x, \xi)$ homogeneous of degree $-k$ in ξ . One applies $\partial_t^2 - \Delta$ to both sides of [47], and obtains an operator of a similar form, with new amplitudes $b^{\pm}(t, x, \xi) \sim \sum b_k^{\pm}(t, x, \xi)$. Setting the terms in this asymptotic expansion equal to zero yields, first for $\varphi^{\pm}(t, x, \xi)$, a partial differential equation known as the eikonal equation:

$$\frac{\partial \varphi^{\pm}}{\partial t} = \pm |\nabla_x \varphi^{\pm}| \tag{49}$$

where $|v|$ is the norm of a vector $v \in T_x M$, determined by the metric tensor. Setting $b_k^{\pm}(t, x, \xi) = 0$ for $k \geq 1$ yields linear differential equations for the amplitude terms in [48], known as transport equations.

Operators of the form [47] are special cases of Fourier integral operators. Seminal works of Keller (1953) and Lax (1957) gave an important stimulus to work on these operators, and work of Hörmander (1971) turned this into a systematic and powerful theory. A particular advance regards

producing a parametrix valid for all t . Generally, one can solve [49] and the associated transport equations for t in some interval, past which the eikonal equation might break down. Hörmander's theory treats products of Fourier integral operators, yielding global constructions. This facilitates the treatment of caustics mentioned earlier. Stationary-phase methods can be brought to bear to relate the singularities of Tb to those of b , when T is a Fourier integral operator.

To construct parametrices for waves reflecting off a boundary, one can again reduce the problem to one of the form [39]. Waves that reflect transversally are given by parametrices of the form [47], although with the role of the variables changed, so that t in [47]–[49] is replaced by a coordinate that vanishes on $\mathbb{R} \times \partial M$.

A parametrix that treats grazing rays can be written in the form of a Fourier–Airy integral operator:

$$u(y) = \int_{\mathbb{R}^n} \left[a A_+(\zeta) + i|\xi|^{-1/3} b A'_+(\zeta) \right] \times A_+(\zeta_0)^{-1} e^{i\theta} \widehat{F}(\xi) d\xi \quad [50]$$

Here $y = (y_1, \dots, y_{n+1})$ denotes a coordinate system on a neighborhood of a boundary point of $\mathbb{R} \times M$, with $y_{n+1} = 0$ on $\mathbb{R} \times \partial M$. We have a pair of phase functions $\theta(y, \xi)$ and $\zeta(y, \xi)$, homogeneous in ξ of degree 1 and $2/3$, respectively, and a pair of amplitudes $a(y, \xi)$ and $b(y, \xi)$, each having asymptotic expansions of the form [48]. The function A_+ is the Airy function [44]. The phase functions satisfy a coupled pair of eikonal equations:

$$\begin{aligned} \langle \nabla_y \theta, \nabla_y \theta \rangle + \zeta \langle \nabla_y \zeta, \nabla_y \zeta \rangle &= 0 \\ \langle \nabla_y \theta, \nabla_y \zeta \rangle &= 0 \end{aligned} \quad [51]$$

where $\langle \cdot, \cdot \rangle$ denotes the Lorentz inner product on $T_y(\mathbb{R} \times M)$ given by $dt^2 - g$. More precisely, [51] is to hold in the region where $\zeta \leq 0$, and also to infinite order at $y_{n+1} = 0$, for $\zeta \geq 0$. One requires $\partial\theta/\partial\xi_j$ to have linearly independent y -gradients, for $j = 1, \dots, n$, and

$$\zeta(y, \xi) = \zeta_0(\xi) = \xi_1^{-1/3} \xi_n \quad \text{for } y_{n+1} = 0 \quad [52]$$

The terms in the asymptotic expansions of $a(y, \xi)$ and $b(y, \xi)$ satisfy coupled systems of transport equations. One can arrange that $b(y, \xi) = 0$ for $y_{n+1} = 0$. Then $u|_{\mathbb{R} \times \partial M} = TF$, where T is a Fourier integral operator, which can be inverted, modulo a smooth error, by Hörmander's theory, producing a parametrix for [39].

The construction of solutions to [51] satisfying [52] is due to Melrose. This followed earlier works of Ludwig (1967), Melrose (1975), and Taylor

(1976), which produced solutions satisfying [52] to infinite order at $\xi_n = 0$. This earlier construction is adequate to produce a grazing ray parametrix, but the sharper result [52] is extremely valuable for constructing a gliding ray parametrix. This has the form

$$u(y) = \int_{\mathbb{R}^n} [a \text{Ai}(\zeta) + i|\xi|^{-1/3} b \text{Ai}'(\zeta)] \times \text{Ai}(\zeta_0)^{-1} e^{i\theta} \widehat{F}(\xi) d\xi \quad [53]$$

It differs from [50] in the use of Ai rather than A_+ . Since Ai has real zeros, it is also convenient to pick $T > 0$ and evaluate θ, ζ, a , and b at $(\xi_1, \dots, \xi_{n-1}, \xi_n + iT)$, and take $\zeta_0 = \xi^{-1/3}(\xi_n + iT)$. The treatment of the eikonal and transport equations is as above, though the Fourier–Airy integral operator [50] has a different behavior from [53], reflecting the difference between how singularities in solutions to the wave equation are carried by grazing and by gliding rays.

Further Reading

- Bowman J, Senior T, and Uslengi P (1969) *Electromagnetic and Acoustic Scattering by Simple Shapes*. Amsterdam: North-Holland.
- Cheeger J and Taylor ME (1982) On the diffraction of waves by conical singularities, I, II. *Communications in Pure and Applied Mathematics* 25: 275–331; 487–529.
- Garnir HG (ed.) (1981) *Singularities in Boundary Value Problems*, NATO Advanced Study Institute Series, vol. 65. Boston: D. Reidel.
- Hörmander L (1971) Fourier integral operators, I. *Acta Mathematica* 127: 79–183.
- Hörmander L (1985) *The Analysis of Partial Differential Operators*, vols. 3–4. New York: Springer.
- Keller JB (1953) The geometrical theory of diffraction. *Proc. Symp. Microwave Optics*, Eaton Electronics Lab, McGill University.
- Lax P (1957) Asymptotic solutions of oscillatory initial value problems. *Duke Mathematical Journal* 24: 627–646.
- Ludwig D (1967) Uniform asymptotic expansion of the field scattered by a convex object at high frequencies. *Communications in Pure and Applied Mathematics* 19: 103–138.
- Melrose RB (1975) Microlocal parametrices for diffractive boundary problems. *Duke Mathematical Journal* 42: 605–635.
- Melrose RB and Sjöstrand J (1978) Singularities of boundary value problems. *Communications in Pure and Applied Mathematics* 31: 593–617.
- Melrose RB and Taylor ME (1985) Near peak scattering and the corrected Kirchhoff approximation for a convex obstacle. *Advances in Mathematics* 55: 242–315.
- Melrose RB and Wunsch J (2004) Propagation of singularities for the wave equation on conic manifolds. *Inventiones Mathematicae* 156: 235–299.
- Taylor ME (1976) Grazing rays and reflection of singularities of solutions to wave equations. *Communications in Pure and Applied Mathematics* 29: 1–38.
- Taylor ME (1981) *Pseudodifferential Operators*. Princeton, NJ: Princeton University Press.
- Taylor ME (1996) *Partial Differential Equations*, vols. 1–3. New York: Springer.

Wavelets: Application to Turbulence

M Farge, Ecole Normale Supérieure, Paris, France

K Schneider, Université de Provence, Marseille, France

© 2006 Published by Elsevier Ltd.

Introduction about Turbulence and Wavelets

What is Turbulence?

Turbulence is a highly nonlinear regime encountered in fluid flows. Such flows are described by continuous fields, for example, velocity or pressure, assuming that the characteristic scale of the fluid motions is much larger than the mean free path of the molecular motions. The prediction of the spacetime evolution of fluid flows from first principles is given by the solutions of the Navier–Stokes equations. The turbulent regime develops when the nonlinear term of Navier–Stokes equations strongly dominates the linear term; the ratio of the norms of both terms is the Reynolds number Re , which characterizes the level of turbulence. In this regime nonlinear instabilities dominate, which leads to the flow sensitivity to initial conditions and unpredictability.

The corresponding turbulent fields are highly fluctuating and their detailed motions cannot be predicted. However, if one assumes some statistical stability of the turbulence regime, averaged quantities, such as mean and variance, or other related quantities, for example, diffusion coefficients, lift or drag, may still be predicted.

When turbulent flows are statistically stationary (in time) or homogeneous (in space), as it is classically supposed, one studies their energy spectrum, given by the modulus of the Fourier transform of the velocity autocorrelation.

Unfortunately, since the Fourier representation spreads the information in physical space among the phases of all Fourier coefficients, the energy spectrum loses all structural information in time or space. This is a major limitation of the classical way of analyzing turbulent flows. This is why we have proposed to use the wavelet representation instead and define new analysis tools that are able to preserve time and space locality.

The same is true for computing turbulent flows. Indeed, the Fourier representation is well suited to study linear motions, for which the superposition principle holds and whose generic behavior is, either to persist at a given scale, or to spread to larger ones. In contrast, the superposition principle does

not hold for nonlinear motions, their archetype being the turbulent regime, which therefore cannot be decomposed into a sum of independent motions that can be separately studied. Generically, their evolution involves a wide range of scales, exciting smaller and smaller ones, even leading to finite-time singularities, e.g., shocks. The “art” of predicting the evolution of such nonlinear phenomena consists of disentangling the active from the passive elements: the former should be deterministically computed, while the latter could either be discarded or their effect statistically modeled. The wavelet representation allows to analyze the dynamics in both space and scale, retaining only those degrees of freedom which are essential to predict the flow evolution. Our goal is to perform a kind of “distillation” and retain only the elements which are essential to compute the nonlinear dynamics.

How One Studies Turbulence?

When studying turbulence one is uneasy about the fact that there are two different descriptions, depending on which side of the Fourier transform one looks from.

- On the one hand, looking from the Fourier space representation, one has a theory which assumes the existence of a nonlinear cascade in an intermediate range of wavenumbers sets, called the “inertial range” where energy is conserved and transferred towards high wavenumbers, but only on average (i.e., considering either ensemble or time or space averages). This implies that a turbulent flow is excited at wavenumbers lower than those of the inertial range and dissipated at wavenumbers higher. Under these hypotheses, the theory predicts that the slope of the energy spectrum in the inertial range scales as $k^{-5/3}$ in dimension 3 and as k^{-3} in dimension 2, k being the wavenumber, i.e., the modulus of the wave vector.
- On the other hand, if one studies turbulence from the physical space representation, there is not yet any universal theory. One relies instead on empirical observations, from both laboratory and numerical experiments, which exhibit the formation and persistence of coherent vortices, even at very high Reynolds numbers. They correspond to the condensation of the vorticity field into some organized structures that contain most of the energy (L^2 -norm of velocity) and enstrophy (L^2 -norm of vorticity).

Moreover, the classical method for modeling turbulent flows consists in neglecting high-wavenumber motions and replacing them by their average, supposing their dynamics to be either linear or slaved to the low wavenumber motions. Such a method would work if there exists a clear separation between low and high wavenumbers, that is, a spectral gap.

Actually, there is now strong evidence, from both laboratory and direct numerical simulation (DNS) experiments, that this is not the case. Conversely, one observes that turbulent flows are nonlinearly active all along the inertial range and that coherent vortices seem to play an essential dynamical role there, especially for transport and mixing. One may then ask the following questions: Are coherent vortices the elementary building blocks of turbulent flows? How can we extract them? Do their mutual interactions have a universal character? Can we compress turbulent flows and compute their evolution with a reduced number of degrees of freedom corresponding to the coherent vortices?

The DNS of turbulent flows, based on the integration of the Navier–Stokes equations using either grid points in physical space or Fourier modes in spectral space, requires a number of degrees of freedom per time step that varies as $Re^{9/4}$ in dimension 3 (and as Re in dimension 2). Due to the inherent limitation of computer performances, one can presently only perform DNS of turbulent flows up to Reynolds numbers $Re = 10^6$. To compute higher Reynolds flows, one should then design *ad hoc* turbulence models, whose parameters are empirically adjusted to each type of flows, in particular to their geometry and boundary conditions, using data from either laboratory or numerical experiments.

What are Wavelets?

The wavelet transform unfolds signals (or fields) into both time (or space) and scale, and possibly directions in dimensions higher than 1. The starting point is a function $\psi \in L^2(\mathbb{R})$, called the “mother wavelet”, which is well localized in physical space $x \in \mathbb{R}$, is oscillating (ψ has at least a vanishing integral, or better, its first m moments vanish), and is smooth (its Fourier transform $\hat{\psi}(k)$ exhibits fast decay for wave numbers $|k|$ tending to infinity). The mother wavelet then generates a family of dilated and translated wavelets

$$\psi_{a,b}(x) = a^{-1/2} \psi\left(\frac{x-b}{a}\right)$$

with $a \in \mathbb{R}^+$ the scale parameter and $b \in \mathbb{R}$ the position parameter, all wavelets being normalized in L^2 -norm.

The wavelet transform of a function $f \in L^2(\mathbb{R})$ is the inner product of f with the analyzing wavelets $\psi_{a,b}$, which gives the wavelet coefficients: $\tilde{f}(a,b) = \langle f, \psi_{a,b} \rangle = \int f(x) \psi_{a,b}(x) dx$. They measure the fluctuations of f around the scale a and the position b . f can then be reconstructed without any loss as the inner product of its wavelet coefficients \tilde{f} with the analyzing wavelets

$$\psi_{a,b} : f(x) = C_\psi^{-1} \iint \tilde{f}(a,b) \psi_{a,b}(x) a^{-2} da db$$

$C_\psi = \int |\hat{\psi}|^2 |k|^{-1} dk$ being a constant which depends on the wavelet ψ .

Like the Fourier transform, the wavelet transform realizes a change of basis from physical space to wavelet space which is an isometry. It thus conserves the inner product (Plancherel theorem), and in particular energy (Parseval’s identity). Let us mention that, due to the localization of wavelets in physical space, the behavior of the signal at infinity does not play any role. Therefore, the wavelet analysis and synthesis can be performed locally, in contrast to the Fourier transform where the nonlocal nature of the trigonometric functions does not allow to perform a local analysis.

Moreover, wavelets constitute building blocks of various function spaces out of which some can be used to construct orthogonal bases. The main difference between the continuous and the orthogonal wavelet transforms is that the latter is non-redundant, but only preserves the invariance by translation and dilation only for a discrete subset of wavelet space which corresponds to the dyadic grid $\lambda = (j, i)$, for which scale is sampled by octaves j and space by positions $2^{-j}i$. The advantage is that all orthogonal wavelet coefficients are decorrelated, which is not the case for the continuous wavelet transform whose coefficients are redundant and correlated in space and scale. Such a correlation can be visualized by plotting the continuous wavelet coefficients of a white noise and the patterns one thus observes are due to the reproducing kernel of the continuous wavelet transform, which corresponds to the correlation between the analyzing wavelets themselves.

In practice, to analyze turbulent signals or fields, one should use the continuous wavelet transform with complex-valued wavelets, since the modulus of the wavelet coefficients allows to read the evolution of the energy density in both space (or time) and scales. If one uses real-valued wavelets instead, the modulus of the wavelet coefficients will present the same oscillations as the analyzing wavelets and it will then become difficult to sort out features

belonging to the signal or to the wavelets. In the case of complex-valued wavelets, the quadrature between the real and the imaginary parts of the wavelet coefficients eliminates these spurious oscillations; this is why we recommend to use complex-valued wavelets, such as the Morlet wavelet. To compress turbulent flows, and *a fortiori* to compute their evolution at a reduced cost, compared to standard methods (finite difference, finite volume, or spectral methods), one should use orthogonal wavelets. This avoids redundancy, since one has the same number of grid points as wavelet coefficients. Moreover there exists a fast algorithm to compute the orthogonal wavelet coefficients which is even faster than the fast Fourier transform, having $O(N)$ operations instead of $O(N \log_2 N)$.

The first paper about the continuous wavelet transform has been published by Grossmann and Morlet (1984). Then, discrete wavelets were constructed, leading to frames (Daubechies *et al.* 1986) and orthogonal bases (Lemarié and Meyer, 1986). From there the formalism of multiresolution analysis (MRA) has been constructed which led to the fast wavelet algorithm (Mallat 1989). The first application of wavelets to analyze turbulent flows has been published by Farge and Rabreau (1988). Since then a long-term research program has been developed for analyzing, computing and modeling turbulent flows using either continuous wavelets, orthogonal wavelets, or wavelet packets.

Wavelet Analysis

Wavelet Spectra

Wavelet space To study turbulent signals one uses the continuous wavelet transform for analysis, and the orthogonal wavelet transform for compression and computation. To perform a continuous wavelet transform, one can choose:

- either a real-valued wavelet, such as the Marr wavelet, also called “Mexican hat,” which is the second derivative of a Gaussian,

$$\psi(x) = (1 - x^2) \exp\left(\frac{-x^2}{2}\right) \quad [1]$$

- or a complex-valued wavelet, such as the Morlet wavelet,

$$\begin{cases} \hat{\psi}(k) = \frac{1}{2\pi} \exp\left(-\frac{(k - k_\psi)^2}{2}\right) & \text{for } k > 0 \\ \hat{\psi}(k) = 0 & \text{for } k \leq 0 \end{cases} \quad [2]$$

with the wavenumber k_ψ denoting the barycenter of the wavelet support in Fourier space computed as

$$k_\psi = \frac{\int_0^\infty k |\hat{\psi}(k)| dk}{\int_0^\infty |\hat{\psi}(k)| dk} \quad [3]$$

For the orthogonal wavelet transform, there is a large collection of possible wavelets and the choice depends on which properties are preferred, for instance: compact support, symmetry, smoothness, number of cancelations, computational efficiency.

From our own experience, we tend to prefer the Coifman wavelet 12, which is compactly supported, has four vanishing moments, is quasi-symmetric, and is defined with a filter of length 12, which leads to a computational cost for the fast wavelet transform in $24N$ operations, since two filters are used.

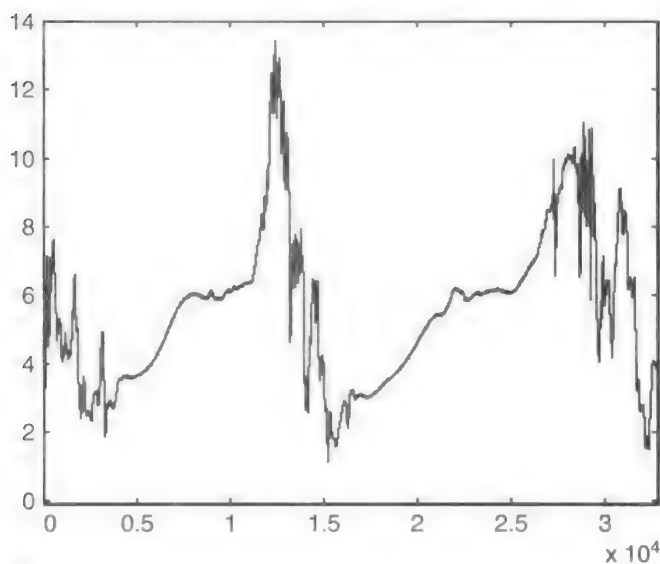
As stated above, we recommend the complex-valued continuous wavelet transform for analysis. In this case, one plots the modulus and the phase of the wavelet coefficients in wavelet space, with a linear horizontal axis for the position b , and a logarithmic vertical axis for the scale a , with the largest scale at the bottom and the smallest scale at the top.

In Figure 1a we show the wavelet analysis of a turbulent signal, corresponding to the time evolution of the velocity fluctuations of two successive vortex breakdowns, measured by hot-wire anemometry at $N = 32768 = 2^{15}$ instants (Cuypers *et al.* 2003). The modulus of the wavelet coefficients (Figure 1b) shows that during the vortex breakdown, which is due to strong nonlinear flow instability, energy is spread over a wide range of scales. The phase of the wavelet coefficients (Figure 1c) is plotted only where the modulus is non-negligible, otherwise the phase information would be meaningless. In Figure 1c, one observes that the lines of constant phase point towards the instants where the signal is less regular, that is, during vortex breakdowns.

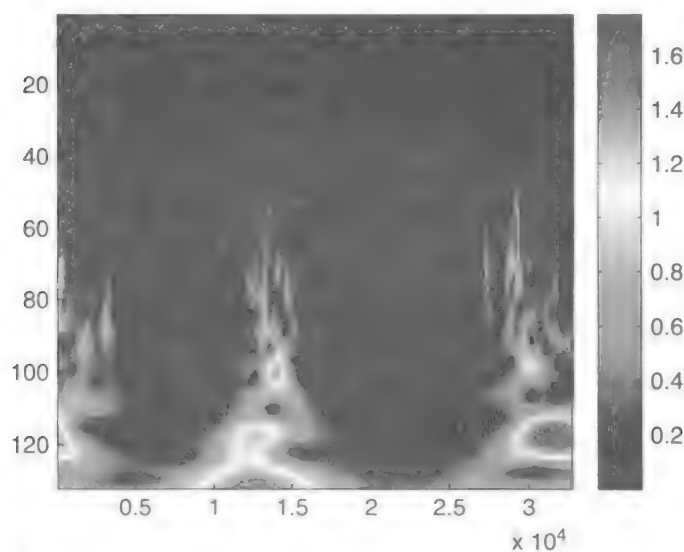
Local wavelet spectrum Since the wavelet transform conserves energy and preserves locality in physical space, one can extend the concept of energy spectrum and define a local energy spectrum, such that

$$\tilde{E}(k, x) = \frac{1}{C_\psi k_\psi} \left| \tilde{f}\left(\frac{k_\psi}{k}, x\right) \right|^2 \quad \text{for } k \geq 0 \quad [4]$$

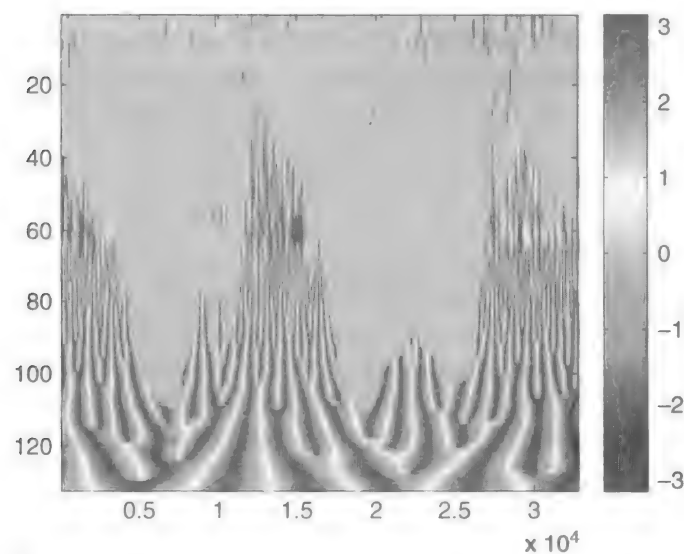
where k_ψ is the centroid wavenumber of the analyzing wavelet ψ and C_ψ is defined in the



(a)



(b)



(c)

Figure 1 Example of a one-dimensional continuous wavelet analysis. (a) the signal to be analyzed, (b) the modulus of its wavelet coefficients, (c) the phase of its wavelet coefficients.

admissibility condition (respectively, eqns [10] and [1] in the article *Wavelets: Mathematical Theory*).

By measuring $\tilde{E}(k, x)$ at different instants or positions, one estimates which elements in the signal contribute most to the global Fourier energy spectrum, in order to suggest a way to decompose the signal into different components. For example, if one considers turbulent flows, one can compare the energy spectrum of the coherent structures (such as isolated vortices in incompressible flows or shocks in compressible flows) and the energy spectrum of the incoherent background flow, since both elements exhibit different correlations and therefore different spectral slopes.

Global wavelet spectrum Although the wavelet transform analyzes the flow using localized functions rather than complex exponentials, one can show that the global wavelet energy spectrum converges towards the Fourier energy spectrum, provided the analyzing wavelet has enough vanishing moments. More precisely, the global wavelet spectrum, defined by integrating [4] over all positions,

$$\tilde{E}(k) = \int_{-\infty}^{\infty} \tilde{E}(k, x) dx \quad [5]$$

gives the correct exponent for a power-law Fourier energy spectrum $E(k) \propto k^{-\beta}$ if the analyzing wavelet has at least $M > (\beta - 1)/2$ vanishing moments. Thus, the steeper the energy spectrum one studies, the more vanishing moments the analyzing wavelet should have.

The inertial range which corresponds to the scales when turbulent flows are dominated by nonlinear interactions, exhibits a power-law behavior as predicted by the statistical theory of homogeneous and isotropic turbulence.

The ability to correctly evaluate the slope of the energy spectrum is an important property of the wavelet transform which is related to its ability to detect and characterize singularities. We will not discuss here how wavelet coefficients could be used to study singularities and fractal measures, since it is presented in detail elsewhere (see *Wavelets: Applications*).

Relation to Classical Analysis

Relation to Fourier spectrum The global wavelet energy spectrum $\tilde{E}(k)$ is actually a smoothed version of the Fourier energy spectrum $E(k)$. This can be

seen from the following relation between the two spectra:

$$\tilde{E}(k) = \frac{1}{C_{\psi} k_{\psi}} \int_0^{\infty} E(k') \left| \hat{\psi} \left(\frac{k_{\psi} k'}{k} \right) \right|^2 dk' \quad [6]$$

which shows that the global wavelet spectrum is an average of the Fourier spectrum weighted by the square of the Fourier transform of the analyzing wavelets at wavenumber k . Note that the larger k , the larger the averaging interval, because wavelets are bandpass filters with $\Delta k/k$ constant. This property of the global wavelet energy spectrum is particularly useful to study turbulent flows. Indeed, the Fourier energy spectrum of a single realization of a turbulent flow is too oscillating to be able to clearly detect a slope, while it is no more the case for the global wavelet energy spectrum, which is a better estimator of the spectral slope.

The real-valued Marr wavelet [1] has only two vanishing moments and thus can correctly measure the energy spectrum exponents up to $\beta < 5$. In the case of the complex-valued Morlet wavelet [2], only the zeroth-order moment is null, but the higher m th order moments are very small ($\propto k_{\psi}^m e^{(-k_{\psi}^2/2)}$), provided that k_{ψ} is larger than 5. For instance, the Morlet wavelet transform with $k_{\psi} = 6$ gives accurate estimates of the power-law exponent of the energy spectrum up to $\beta < 7$.

There is also a family of wavelets with an infinite number of cancelations

$$\hat{\psi}_n(k) = \alpha_n \exp \left(-\frac{1}{2} \left(k^2 + \frac{1}{k^{2n}} \right) \right) \quad n \geq 1 \quad [7]$$

where α_n is chosen for normalization.

These wavelets can therefore correctly measure any power-law energy spectrum, and thus detect the difference between a power-law energy spectrum and a Gaussian energy spectrum ($E(k) \propto e^{-(k/k_0)^2}$). For instance, it is important in turbulence to determine the wavenumber after which the energy spectrum decays exponentially, since this wavenumber defines the end of the inertial range, dominated by nonlinear interactions, and the beginning of the dissipative range, dominated by linear dissipation.

Relation to structure functions In this subsection we will point out the limitations of classical measures of intermittency and present a set of wavelet-based alternatives.

The classical measures based on structure functions can be thought of as a special case of wavelet filtering using a nonsmooth wavelet defined as the difference of two Diracs (DOD). It is this lack of regularity of the underlying wavelet that limits the adequacy of classical measures to analyze smooth signals. Wavelet-based diagnostics can overcome these limitations, and produce accurate results, whatever the signal to be analyzed.

We will link the scale-dependent moments of the wavelet coefficients and the structure functions, which are classically used to study turbulence. In the case of second-order statistics, the global wavelet spectrum corresponds to the second-order structure function. Furthermore, a rigorous bound for the maximum exponent detected by the structure functions can be computed, but there is a way to overcome this limitation by using wavelets.

The increments of a signal, also called the modulus of continuity, can be seen as its wavelet coefficients using the DOD wavelet

$$\psi^{\delta}(x) = \delta(x+1) - \delta(x) \quad [8]$$

We thus obtain

$$f(x+a) - f(x) = \tilde{f}_{x,a} = \langle f, \psi_{x,a}^{\delta} \rangle \quad [9]$$

with $\psi_{x,a}(y) = 1/a [\delta((y-x)/a + 1) - \delta((y-x)/a)]$. Note that the wavelet is normalized with respect to the L^1 -norm. The p th-order structure function $S_p(a)$ therefore corresponds to the p th-order moment of the wavelet coefficients at scale a

$$S_p(a) = \int (\tilde{f}_{x,a})^p dx \quad [10]$$

As the DOD wavelet has only one vanishing moment (its mean), the exponent of the p th-order structure function in the case of a self-similar behavior is limited by p , that is, if $S_p(a) \propto a^{\zeta(p)}$, then $\zeta(p) < p$. To be able to detect larger exponents, one has to use increments with a larger stencil, or wavelets with more vanishing moments.

We now concentrate on the case $p=2$, that is, the energy norm. Equation [6] gives the relation between the global wavelet spectrum $\tilde{E}(k)$ and the Fourier spectrum $E(k)$ for an arbitrary wavelet ψ . For the DOD wavelet we find, since $\hat{\psi}^{\delta}(k) = e^{ik} - 1 = e^{ik/2} (e^{ik/2} - e^{-ik/2})$ and hence $|\hat{\psi}^{\delta}(k)|^2 = 2(1 - \cos k)$, that

$$\tilde{E}(k) = \frac{1}{C_{\psi} k} \int_0^{\infty} E(k') \left(2 - 2 \cos \left(\frac{k_{\psi} k'}{k} \right) \right) dk' \quad [11]$$

Setting $a = k_\psi/k$, we see that the wavelet spectrum corresponds to the second-order structure function, such that

$$\tilde{E}(k) = \frac{1}{C_\psi k} S_2(a) \quad [12]$$

The above results show that, if the Fourier spectrum behaves like $k^{-\alpha}$ for $k \rightarrow \infty$, $\tilde{E}(k) \propto k^{-\alpha}$ if $\alpha < 2M + 1$, where M denotes the number of vanishing moments of the wavelets. Consequently, we find for $S_2(a)$ that $S_2(a) \propto a^{\zeta(p)} = (k_\psi/k)^{\zeta(p)}$ for $a \rightarrow 0$ if $\zeta(2) \leq 2M$. For the DOD wavelet, we have $M = 1$, therefore, the second-order structure function can only detect slopes smaller than 2, corresponding to an energy spectrum whose slope is shallower than -3 . Thus, the usual structure functions give spurious results for sufficiently smooth signals. The relation between structure functions and wavelet coefficients can be generalized in the context of Besov spaces, which are classically used for non-linear approximation theory (see Wavelets: Mathematical Theory).

Intermittency Measures

Intermittency is defined as localized bursts of high-frequency activity. This means that intermittent phenomena are localized in both physical and spectral spaces, and thus a suitable basis for representing intermittency should reflect this dual localization. The Fourier basis is well localized in spectral space, but delocalized in physical space. Therefore, when a turbulence signal is filtered using a high-pass Fourier transform and then reconstructed in physical space, for example, to calculate the flatness, some spatial information is lost. This leads to smoothing of strong gradients and spurious oscillations in the background, which come from the fact that the modulus and phase of the discarded high wavenumber Fourier modes have been lost. The spatial errors introduced by such a Fourier filtering lead to errors in estimating the flatness, and hence the signal's intermittency.

When a quantity (e.g., velocity derivative) is intermittent, it contains rare but strong events (i.e., bursts of intense activity), which correspond to large deviations reflected in the "heavy tails" of the PDF. Second-order statistics (e.g., energy spectrum, second-order structure function) are relatively insensitive to such rare events whose time or space supports are very small and thus do not dominate the integral. However, these events become increasingly important for higher-order statistics, where they finally dominate. High-order

statistics therefore characterize intermittency. Of course, intermittency is not essential for all problems: second-order statistics are sufficient to measure dispersion (dominated by energy-containing scales), but not to calculate drag or mixing (dominated by vorticity production in thin boundary or shear layers).

To measure intermittency, one uses the space-scale information contained in the wavelet coefficients to define scale-dependent moments and moment ratios. Useful diagnostics to quantify the intermittency of a field f are the moments of its wavelet coefficients at different scales j

$$M_{p,j}(f) = 2^{-j} \sum_{i=0}^{2^j-1} |\tilde{f}_{j,i}|^p \quad [13]$$

Note that the distribution of energy scale by scale, that is, the scalogram, can be computed from the second-order moment of the orthogonal wavelet coefficients: $E_j = 2^{j-1} M_{2,j}$. Due to orthogonality of the decomposition, the total energy is just the sum: $E = \sum_{j \geq 0} E_j$.

The sparsity of the wavelet coefficients at each scale is a measure of intermittency, and it can be quantified using ratios of moments at different scales

$$Q_{p,q,j}(f) = \frac{M_{p,j}(f)}{(M_{q,j}(f))^{p/q}} \quad [14]$$

which may be interpreted as quotient norms computed in two different functional spaces, L^p - and L^q -spaces. Classically, one chooses $q = 2$ to define typical statistical quantities as a function of scale. Recall that for $p = 4$ we obtain the scale-dependent flatness $F_j = Q_{4,2,j}$. It is equal to 3 for a Gaussian white noise at all scales j , which proves that this signal is not intermittent. The scale-dependent skewness, hyperflatness, and hyperskewness are obtained for $p = 3, 5$, and 6 , respectively. For intermittent signals $Q_{p,q,j}$ increases with j , whatever p and q .

Wavelet Compression

Principle

To study turbulent signals, we now propose to separate the rare and extreme events from the dense events, and then calculate their statistics independently. A major difficulty in turbulence research is that there is no clear scale separation between these two kinds of events. This lack of "spectral gap" excludes Fourier filtering for disentangling these two behaviors. Since the rare events are well

localized in physical space, one might try to use an on-off filter defined in physical space to extract them. However, this approach changes the spectral properties by introducing spurious discontinuities, adding an artificial scaling (e.g., k^{-2} in one dimension) to the energy spectrum. To avoid these problems, we use the wavelet representation, which combines both physical and spectral space localizations (bounded from below by Heisenberg's uncertainty principle). In turbulence, the relevant rare events are the coherent vortices and the dense events correspond to the residual background flow. We have proposed a nonlinear wavelet filtering of the wavelet coefficients of vorticity to extract the coherent vortices out of turbulent flows. We now detail the different steps of this procedure.

Extraction of Coherent Structures

Principle We propose a new method to extract coherent structures from turbulent flows, as encountered in fluids (e.g., vortices, shocklets) or plasmas (e.g., bursts), in order to study their role in transport and mixing.

We first replace the Fourier representation by the wavelet representation, which keeps track of both time and scale, instead of frequency only. The second improvement consists in changing our viewpoint about coherent structures. Since there is not yet a universal definition of coherent structures, we prefer starting from a minimal but more consensual statement about them, that everyone hopefully could agree with: "coherent structures are not noise." Using this apophatic method, we propose the following definition: "coherent structures are what remain after denoising."

For the noise we use the mathematical definition stating that a noise cannot be compressed in any functional basis. Another way to say this is to observe that the shortest description of a noise is the noise itself. Notice that often one calls "noise" what is actually "experimental noise," but not noise in the mathematical sense.

Considering our definition of coherent structures, turbulent signals can be split into two contributions: coherent bursts, corresponding to that part of the signal which can be compressed in a wavelet basis, and incoherent noise, corresponding to that part of the signal which cannot be compressed, neither in wavelets nor in any other basis. We will then check *a posteriori* that the incoherent contribution is spread, and therefore does not compress, in both Fourier and grid-point basis. Since we use the orthogonal wavelet representation, both coherent and incoherent components are

orthogonal and therefore the L^2 -norm, for example, energy or enstrophy, is a superposition of coherent and incoherent contributions (Mallat 1998).

Assuming that coherent structures are what remain after denoising, we need a model, not for the structures themselves, but for the noise. As a first guess, we choose the simplest model and suppose the noise to be additive, Gaussian and white, that is, uncorrelated. Having this model in mind, we use Donoho and Johnstone's theorem to compute the value to threshold the wavelet coefficients. Since the threshold value depends on the variance of the noise, which in the case of turbulence is not *a priori* known, we propose a recursive method to estimate it from the variance of the weakest wavelet coefficients, that is, those whose modulus is below the threshold value.

Wavelet decomposition We describe the wavelet algorithm to extract coherent vortices out of turbulent flows and apply it as example to a 3D turbulent flow. We consider the vorticity field $\omega = \nabla \times \mathbf{v}$, computed at resolution $N = 2^{3J}$, N being the number of grid points and J the number of octaves in each spatial direction. Each vorticity component is developed into an orthogonal wavelet series from the largest scale $l_{\max} = 2^0$ to the smallest scale $l_{\min} = 2^{J-1}$ using a three-dimensional (3D) MRA:

$$\omega(\mathbf{x}) = \bar{\omega}_{0,0,0} \phi_{0,0,0}(\mathbf{x}) + \sum_{j=0}^{J-1} \sum_{i_x=0}^{2^j-1} \sum_{i_y=0}^{2^j-1} \sum_{i_z=0}^{2^j-1} \sum_{d=1}^7 \bar{\omega}_{j,i_x,i_y,i_z}^d \psi_{j,i_x,i_y,i_z}^d(\mathbf{x}) \quad [15]$$

with $\phi_{j,i_x,i_y,i_z}(\mathbf{x}) = \phi_{j,i_x}(x)\phi_{j,i_y}(y)\phi_{j,i_z}(z)$, and

$$\psi_{j,i_x,i_y,i_z}^d(\mathbf{x}) = \begin{cases} \psi_{j,i_x}(x)\phi_{j,i_y}(y)\phi_{j,i_z}(z) & d=1 \\ \phi_{j,i_x}(x)\psi_{j,i_y}(y)\phi_{j,i_z}(z) & d=2 \\ \phi_{j,i_x}(x)\phi_{j,i_y}(y)\psi_{j,i_z}(z) & d=3 \\ \psi_{j,i_x}(x)\phi_{j,i_y}(y)\psi_{j,i_z}(z) & d=4 \\ \psi_{j,i_x}(x)\psi_{j,i_y}(y)\phi_{j,i_z}(z) & d=5 \\ \phi_{j,i_x}(x)\psi_{j,i_y}(y)\psi_{j,i_z}(z) & d=6 \\ \psi_{j,i_x}(x)\psi_{j,i_y}(y)\psi_{j,i_z}(z) & d=7 \end{cases} \quad [16]$$

where $\phi_{j,i}$ and $\psi_{j,i}$ are the one-dimensional scaling function and the corresponding wavelet, respectively. Due to orthogonality, the scaling coefficients are given by $\bar{\omega}_{0,0,0} = \langle \omega, \phi_{0,0,0} \rangle$ and the wavelet coefficients are given by $\bar{\omega}_{j,i_x,i_y,i_z}^d = \langle \omega, \psi_{j,i_x,i_y,i_z}^d \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the L^2 -inner product.

Nonlinear thresholding The vorticity field is then split into ω_C and ω_I by applying a nonlinear thresholding to the wavelet coefficients. The threshold is defined as $\epsilon = (\frac{4}{3} Z \ln N)^{1/2}$. It only depends on the total enstrophy $Z = \frac{1}{2} \int |\omega|^2 dx$ and on the number of grid points N without any adjustable parameter. The choice of this threshold is based on theorems by Donoho and Johnstone proving optimality of the wavelet representation to denoise signals in the presence of Gaussian white noise, since this wavelet-based estimator minimizes the maximal L^2 -error for functions with inhomogeneous regularity (Mallat 1998).

Wavelet reconstruction The coherent vorticity field ω_C is reconstructed from the wavelet coefficients whose modulus is larger than ϵ and the incoherent vorticity field ω_I from the wavelet coefficients whose modulus is smaller or equal to ϵ . The two fields thus obtained, ω_C and ω_I , are orthogonal, which ensures a separation of the total enstrophy into $Z = Z_C + Z_I$ because the interaction term $\langle \omega_C, \omega_I \rangle$ vanishes. We then use Biot–Savart’s relation $v = \nabla \times (\nabla^{-2} \omega)$ to reconstruct the coherent velocity v_C and the incoherent velocity v_I from the coherent and incoherent vorticities, respectively.

Application to 3D Turbulence

We consider a 3D homogeneous isotropic turbulent flow, computed by DNS at resolution $N = 256^3$, which corresponds to a Reynolds number based on the Taylor microscale $R_\lambda = 168$ (Farge *et al.* 2003). The computation uses a pseudospectral code, with a Gaussian random vorticity field as initial condition, and the flow evolution is integrated until a statistically stationary state is reached. Figure 2 shows the modulus of the vorticity fluctuations of the total flow, zooming on a 64^3 subcube to enhance structural details. The flow exhibits elongated, distorted, and folded vortex tubes, as observed in laboratory and numerical experiments.

We apply to the total flow the wavelet compression algorithm described above. We find that only 2.9% wavelet modes correspond to the coherent flow, which retains 79% of the energy (L^2 -norm of velocity) and 75% of the enstrophy (L^2 -norm of vorticity), while the remaining 97.1% incoherent modes contain only 1% of the energy and 21% of the enstrophy. We display the modulus of the coherent (Figure 3) and incoherent (Figure 4) vorticity fluctuations resulting from the wavelet decomposition.

Note that the values of the three isosurfaces chosen for visualization ($|\omega| = 6Z^{1/2}$, $8Z^{1/2}$ and $10Z^{1/2}$, with Z the total enstrophy) are the same for the total and

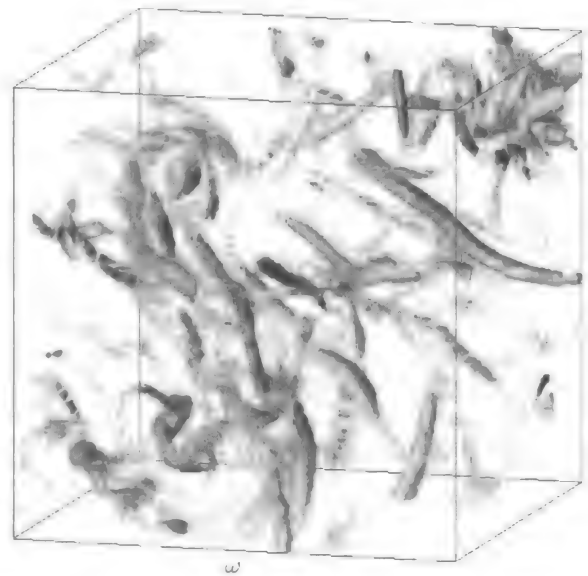


Figure 2 Isosurfaces of total vorticity field, for $|\omega| = 3\sigma, 4\sigma, 5\sigma$ with opacity 1, 0.5, 0.1, respectively, and σ^2 the total enstrophy. Simulation with resolution $N = 256^3$ for $R_\lambda = 168$. Zoom on a subcube 64^3 . Reprinted with permission from Farge *et al.* Coherent vortex extraction in three-dimensional homogeneous turbulence: Comparison between CVS-wavelet and POD-Fourier decompositions. *Physics of Fluids* 15(10): 2886–2896. Copyright 2003, American Institute of Physics.

coherent vorticities, but they have been reduced by a factor 2 for the incoherent vorticity whose fluctuations are much smaller. In the coherent vorticity (Figure 3) we recognize the same vortex tubes as those present in the total vorticity (Figure 2). In contrast, the remaining vorticity (Figure 4) is much more homogeneous and

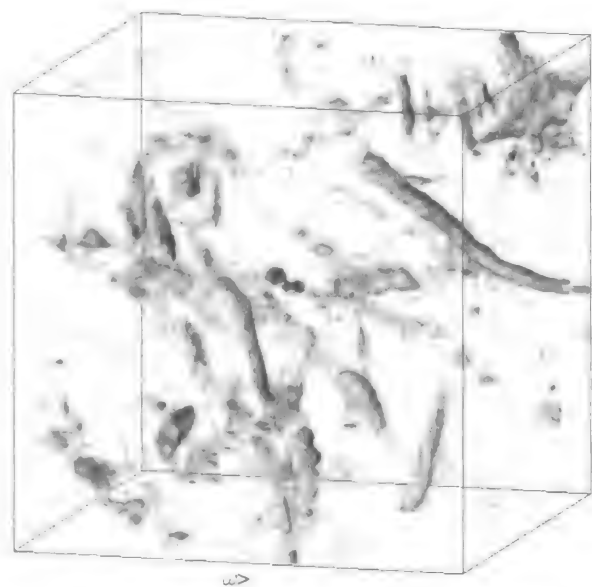


Figure 3 Isosurfaces of coherent vorticity field, for $|\omega| = 3\sigma, 4\sigma, 5\sigma$ with opacity 1, 0.5, 0.1, respectively. Simulation with resolution $N = 256^3$. Zoom on a subcube 64^3 . Reprinted with permission from Farge *et al.* Coherent vortex extraction in three-dimensional homogeneous turbulence: Comparison between CVS-wavelet and POD-Fourier decompositions. *Physics of Fluids* 15(10): 2886–2896. Copyright 2003, American Institute of Physics.

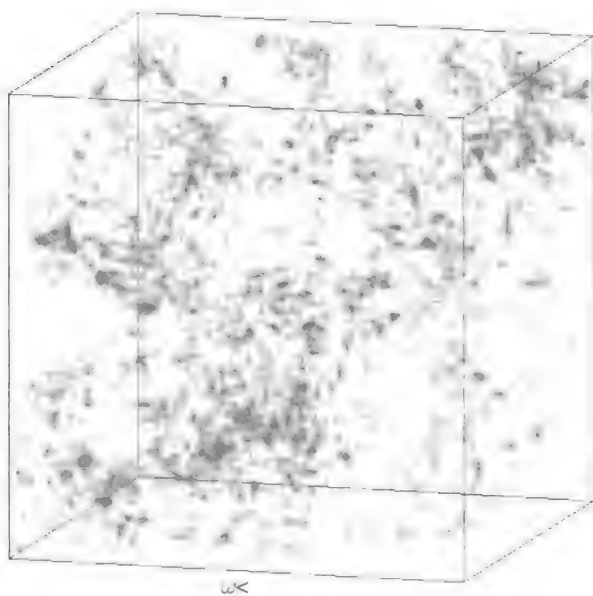


Figure 4 Isosurfaces of incoherent vorticity field, for $|\omega| = 3/2\sigma, 2\sigma, 5/2\sigma$ with opacity 1, 0.5, 0.1, respectively. Simulation with resolution $N = 256^3$. Zoom on a subcube 64^3 . Reprinted with permission from Farge *et al.* Coherent vortex extraction in three-dimensional homogeneous turbulence: Comparison between CVS-wavelet and POD-Fourier decompositions. *Physics of Fluids* 15(10): 2886–2896. Copyright 2003, American Institute of Physics.

does not exhibit coherent structures. Hence, the wavelet compression retains all the vortex tubes and preserves their structure at all scales. Consequently, the coherent flow is as intermittent as the total flow, while the incoherent flow is structureless and non intermittent. Modeling the effect of the incoherent flow onto the coherent flow should then be much simpler than with methods based on Fourier filtering.

Figure 5 shows the velocity PDF in semilogarithmic coordinates. We observe that the coherent velocity has

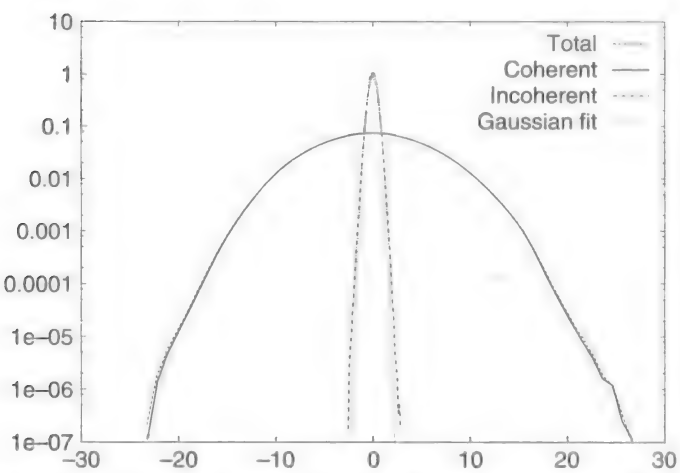


Figure 5 Velocity PDF, resolution $N = 256^3$ with a zoom at 64^3 . Reprinted with permission from Farge *et al.* Coherent vortex extraction in three-dimensional homogeneous turbulence: Comparison between CVS-wavelet and POD-Fourier decompositions. *Physics of Fluids* 15(10): 2886–2896. Copyright 2003, American Institute of Physics.

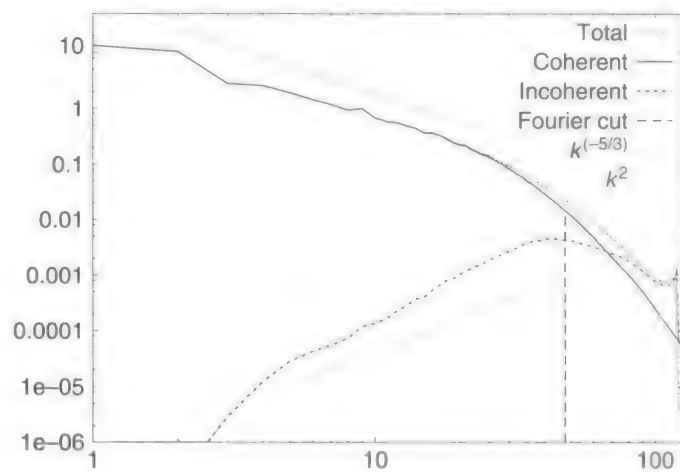


Figure 6 Energy spectrum, resolution $N = 256^3$ with a zoom at 64^3 . Reprinted with permission from Farge *et al.* Coherent vortex extraction in three-dimensional homogeneous turbulence: Comparison between CVS-wavelet and POD-Fourier decompositions. *Physics of Fluids* 15(10): 2886–2896. Copyright 2003, American Institute of Physics.

the same Gaussian distribution as the total velocity, while the incoherent velocity remains Gaussian, but its variance is much smaller. The corresponding energy spectra are plotted on Figure 6. We observe that the spectrum of the coherent energy is identical to the spectrum of the total energy all along the inertial range. This implies that the vortex tubes are responsible for the $k^{-5/3}$ energy scaling, which corresponds to a long-range correlation, characteristic of 3D turbulence as predicted by Kolmogorov’s theory. In contrast, the incoherent energy has a scaling close to k^2 , which corresponds to an energy equipartition between all wave vectors k , since the isotropic spectrum is obtained by integrating energy in 3D k -space over 2D shells $k = |k|$. The incoherent velocity field is therefore spatially uncorrelated, which is consistent with the observation that incoherent vorticity is structureless and homogeneous.

From these observations, we propose the following scenario to interpret the turbulent cascade: the coherent energy injected at large scales is transferred towards small scales by nonlinear interactions between vortex tubes. In the meantime, these nonlinear interactions also produce incoherent energy at all scales, which is dissipated at the smallest scales by molecular kinematic viscosity. Thus, the coherent flow causes direct transfer of the coherent energy into incoherent energy. Conversely, the incoherent flow does not trigger any energy transfer to the coherent flow, as it is structureless and uncorrelated. We conjecture that the coherent flow is dynamically active, while the incoherent flow is slaved to it, being only passively advected and mixed by the coherent vortex tubes. This is a different view from the classical interpretation since it does not suppose any scale separation. Both

coherent and incoherent flows are active all along the inertial range, but they are characterized by different probability distribution functions and correlations: non-Gaussian and long-range correlated for the former, while Gaussian and uncorrelated for the latter.

Wavelet Computation

Principle

The mathematical properties of wavelets (*see* Wavelets: Mathematical Theory) motivate their use for solving of partial differential equations (PDEs).

The localization of wavelets, both in scale and space, leads to effective sparse representations of functions and pseudodifferential operators (and their inverse) by performing nonlinear thresholding of the wavelet coefficients of the function and of the matrices representing the operators. Wavelet coefficients allow to estimate the local regularity of solutions of PDEs and thus can define autoadaptive discretizations with local mesh refinements. The characterization of function spaces in terms of wavelet coefficients and the corresponding norm equivalences lead to diagonal preconditioning of operators in wavelet space.

Moreover, the existence of the fast wavelet transform yields algorithms with optimal linear complexity. The currently existing algorithms can be classified in different ways. We can distinguish between Galerkin, collocation, and hybrid schemes. Hybrid schemes combine classical discretizations, for example, finite differences or finite volumes, and wavelets, which are only used to speed up the linear algebra and to define adaptive grids. On the other hand, Galerkin and collocation schemes employ wavelets directly for the discretization of the solution and the operators. Wavelet methods have been developed to solve Burger's, Stokes, Kuramoto–Sivashinsky, nonlinear Schrödinger, Euler, and Navier–Stokes equations. As an example, we present an adaptive wavelet algorithm, of Galerkin type, to solve the 2D Navier–Stokes equations.

Adaptive Wavelet Scheme

We consider the 2D Navier–Stokes equations written in terms of vorticity ω and stream function Ψ , which are both scalars in two dimensions,

$$\partial_t \omega + \mathbf{v} \cdot \nabla \omega - \nu \nabla^2 \omega = \nabla \times \mathbf{F} \quad [17]$$

$$\nabla^2 \Psi = \omega \quad \text{and} \quad \mathbf{v} = \nabla^\perp \Psi \quad [18]$$

for $\mathbf{x} \in [0, 1]^2$, $t > 0$. The velocity is denoted by \mathbf{v} , \mathbf{F} is an external force, $\nu > 0$ is the molecular kinematic viscosity, and $\nabla^\perp = (-\partial_y, \partial_x)$.

The above equations are completed with boundary conditions and a suitable initial condition.

Time discretization Introducing a classical semi-implicit time discretization with a time step Δt and setting $\omega^n(\mathbf{x}) \approx \omega(\mathbf{x}, n\Delta t)$, we obtain

$$(1 - \nu \Delta t \nabla^2) \omega^{n+1} = \omega^n + \Delta t (\nabla \times \mathbf{F}^n - \mathbf{v}^n \cdot \nabla \omega^n) \quad [19]$$

$$\nabla^2 \Psi^{n+1} = \omega^{n+1} \quad \text{and} \quad \mathbf{v}^{n+1} = \nabla^\perp \Psi^{n+1} \quad [20]$$

Hence, in each time step two elliptic problems have to be solved and a differential operator has to be applied.

Formally the above equations can be written in the abstract form $L\mathbf{u} = \mathbf{f}$, where L is an elliptic operator with constant coefficients. This corresponds to a Helmholtz type equation for ω with $L = (1 - \nu \Delta t \nabla^2)$ and a Poisson equation for Ψ with $L = \nabla^2$.

Spatial discretization For the spatial discretization, we use the method of weighted residuals, that is, a Petrov–Galerkin scheme. The trial functions are orthogonal wavelets ϕ and the test functions are operator adapted wavelets, called “vaguelettes,” θ . To solve the elliptic equation $L\mathbf{u} = \mathbf{f}$ at time step t^{n+1} , we develop \mathbf{u}^{n+1} into an orthogonal wavelet series, that is, $\mathbf{u}^{n+1} = \sum_\lambda \tilde{u}_\lambda^{n+1} \psi_\lambda$, where $\lambda = (j, i_x, i_y, d)$ denotes the multi-index for scale j , space i , and direction d . Requiring that the residual vanishes with respect to all test functions θ_λ , we obtain a linear system for the unknown wavelet coefficients \tilde{u}_λ^{n+1} of the solution \mathbf{u} :

$$\sum_\lambda \tilde{u}_\lambda^{n+1} \langle L\psi_\lambda, \theta_{\lambda'} \rangle = \langle \mathbf{f}, \theta_{\lambda'} \rangle \quad [21]$$

The test functions θ are defined such that the stiffness matrix turns out to be the identity. Therefore, the solution of $L\mathbf{u} = \mathbf{f}$ reduces to a change of basis, that is, $\mathbf{u}^{n+1} = \sum_\lambda \langle \mathbf{f}, \theta_\lambda \rangle \psi_\lambda$. The right-hand side (RHS) \mathbf{f} can then be developed into a biorthogonal operator adapted wavelet basis $\mathbf{f} = \sum_\lambda \langle \mathbf{f}, \theta_\lambda \rangle \zeta_\lambda$, with $\theta_\lambda = L^{*-1} \psi_\lambda$ and $\zeta_\lambda = L\psi_\lambda$, $*$ denoting the adjoint operator. By construction, θ and ζ are biorthogonal, that is, such that $\langle \theta_\lambda, \zeta_{\lambda'} \rangle = \delta_{\lambda, \lambda'}$. It can be shown that both have similar localization properties in physical and Fourier space as ψ , and that they form a Riesz basis.

Adaptive discretization To get an adaptive space discretization for the linear problem $L\mathbf{u} = \mathbf{f}$, we

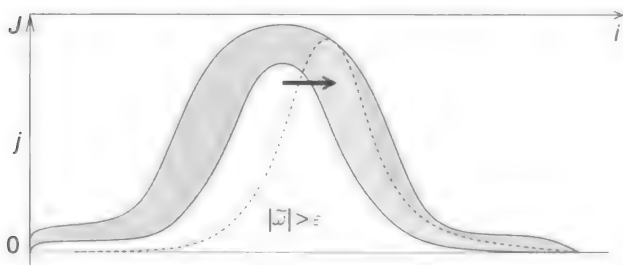


Figure 7 Illustration of the dynamic adaption strategy in wavelet coefficient space.

consider only the significant wavelet coefficients of the solution. Hence, we only retain coefficients \tilde{u}_λ^n whose modulus is larger than a given threshold ε , that is, $|\tilde{u}_\lambda^n| > \varepsilon$. The corresponding coefficients are shown in **Figure 7** (white area under the solid line curve).

Adaption strategy To be able to integrate the equation in time we have to account for the evolution of the solution in wavelet coefficient space (indicated by the arrow in **Figure 7**). Therefore, we add at time step t^n the neighbors to the retained coefficients, which constitute a security zone (gray area in **Figure 7**). The equation is then solved in this enlarged coefficient set (white and gray areas below the curves in **Figure 7**) to obtain \tilde{u}_λ^{n+1} . Subsequently, we threshold the coefficients and retain only those whose modulus $|\tilde{u}_\lambda^{n+1}| > \varepsilon$ (coefficients under the dashed curve in **Figure 7**). This strategy is applied in each time step and hence allows to automatically track the evolution of the solution in both scale and space.

Evaluation of the nonlinear term For the evaluation of the nonlinear term $f(u^n)$, where the wavelet coefficients \tilde{u}^n are given, there are two possibilities:

- *Evaluation in wavelet coefficient space.* As illustration, we consider a quadratic nonlinear term, $f(u) = u^2$. The wavelet coefficients of f can be calculated using the connection coefficients, that is, one has to calculate the bilinear expression, $\sum_\lambda \sum_{\lambda'} \tilde{u}_\lambda \mathcal{I}_{\lambda\lambda'\lambda''} \tilde{u}_{\lambda'}$ with the interaction tensor $\mathcal{I}_{\lambda\lambda'\lambda''} = \langle \psi_\lambda \psi_{\lambda'}, \theta_{\lambda''} \rangle$. Although many coefficients of \mathcal{I} are zero or very small, the size of \mathcal{I} leads to a computation which is quite untractable in practice.
- *Evaluation in physical space.* This approach is similar to the pseudospectral evaluation of the nonlinear terms used in spectral methods, therefore it is called pseudowavelet technique. The

advantage of this scheme is that general nonlinear terms, for example, $f(u) = (1 - u)e^{-C/u}$, can be treated more easily. The method can be summarized as follows: starting from the significant wavelet coefficients, $|\tilde{u}_\lambda| > \varepsilon$, one reconstructs u on a locally refined grid and gets $u(x_\lambda)$. Then one can evaluate $f(u(x_\lambda))$ pointwise and the wavelet coefficients \tilde{f}_λ are calculated using the adaptive decomposition.

Finally, one computes the scalar products of the RHS of [21] with the test functions θ to advance the solution in time. We compute $\tilde{u}_\lambda = \langle f, \theta_\lambda \rangle$ belonging to the enlarged coefficient set (white and gray regions in **Figure 7**).

The algorithm is of $O(N)$ complexity, where N denotes the number of wavelet coefficients retained in the computation.

Application to 2D Turbulence

To illustrate the above algorithm we present an adaptive wavelet computation of a vortex dipole in a square domain, impinging on a no-slip wall at Reynolds number $Re = 1000$. To take into account the solid wall, we use a volume penalization method, for which both the fluid flow and the solid container are modeled as a porous medium whose porosity tends towards zero in the fluid and towards infinity in the solid region.

The 2D Navier–Stokes equations are thus modified by adding the forcing term $F = -(1/\eta)\chi v$ in eqn [18], where η is the penalization parameter and χ is the characteristic function whose value is 1 in the solid region and 0 elsewhere. The equations are solved using the adaptive wavelet method in a periodic square domain of size 1.1, in which the square container of size 1 is imbedded, taking $\eta = 10^{-3}$. The maximal resolution corresponds to a fine grid of 1024^2 points. **Figure 8a** shows snapshots of the vorticity field at times $t = 0.2, 0.4, 0.6$, and 0.8 (in arbitrary units). We observe that the vortex dipole is moving towards the wall and that strong vorticity gradients are produced when the dipole hits the wall. The computational grid is dynamically adapted during the flow evolution, since the nonlinear wavelet filter automatically refines the grid in regions where strong gradients develop. **Figure 8b** shows the centers of the retained wavelet coefficients at corresponding times.

Note that during the computation only 5% out of 1024^2 wavelet coefficients are used. The time evolution of total kinetic energy and the total enstrophy $F = -(\frac{1}{\eta})\chi v$, are plotted in **Figure 9** to

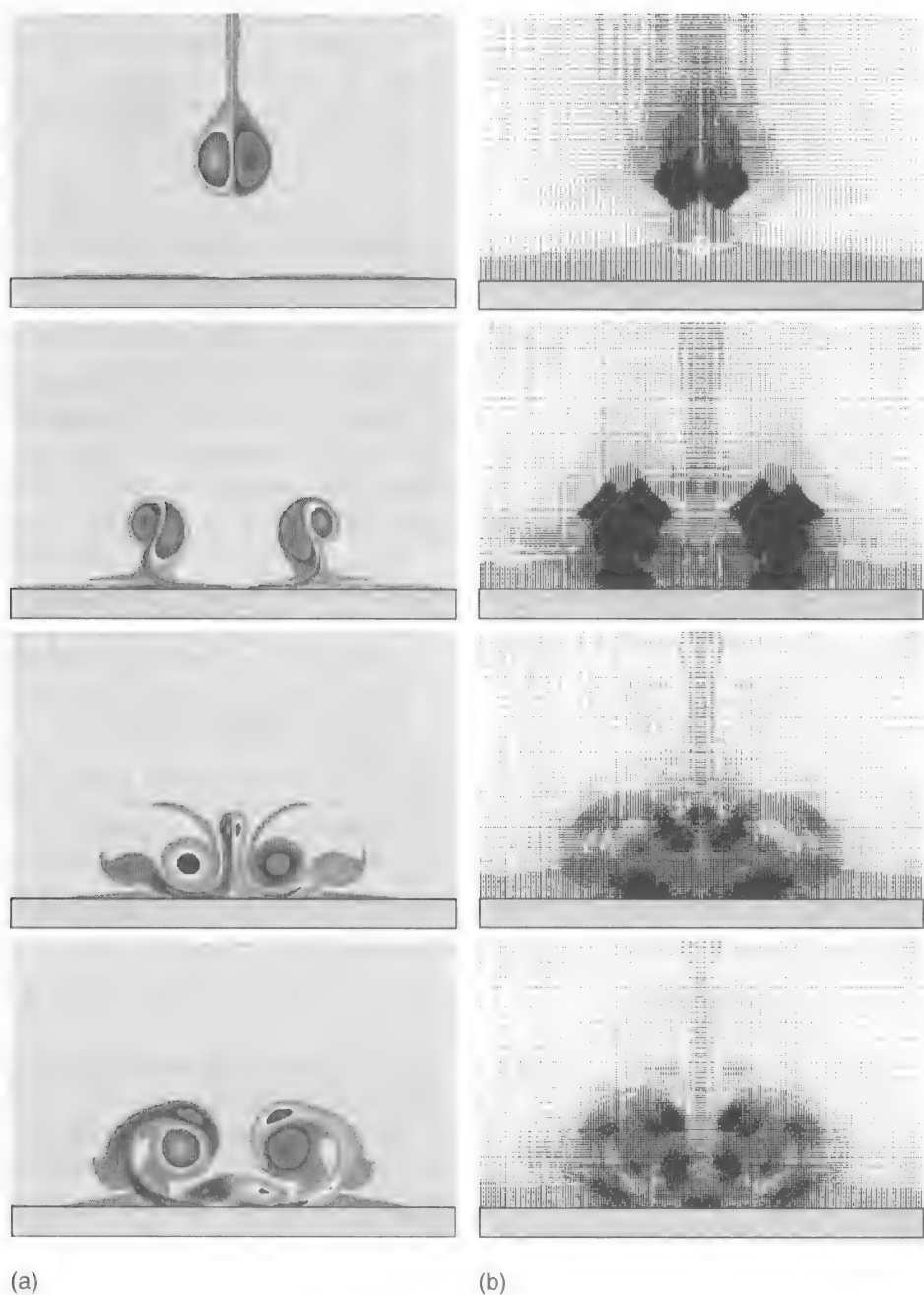


Figure 8 Dipole wall interaction at $Re = 1000$. (a) Vorticity field, (b) corresponding centers of the active wavelets, at $t = 0.2, 0.4, 0.6$, and 0.8 (from top to bottom).

show the production of enstrophy and the concomitant dissipation of energy when the vortex dipole hits the wall.

This computation illustrates the fact that the adaptive wavelet method allows an automatic grid refinement, both in the boundary layers at the wall and also in shear layers which develop during the flow evolution far from the wall. Therewith, the number of grid points necessary for the computation is significantly reduced, and we conjecture that the resulting compression rate will increase with the Reynolds number.

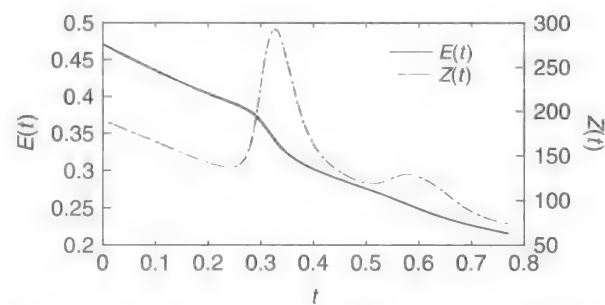


Figure 9 Time evolution of energy (solid line) and enstrophy (dashed line).

Acknowledgments

Marie Farge thankfully acknowledges Trinity College, Cambridge, UK, and the Centre International de Rencontres Mathématiques (CIRM), Marseille, France, for hospitality while writing this paper.

See also: Turbulence Theories; Viscous Incompressible Fluids: Mathematical Theory; Wavelets: Applications; Wavelets: Mathematical Theory.

Further Reading

- Cohen A (2000) Wavelet methods in numerical analysis. In: Ciarlet PG and Lions JL (eds.) *Handbook of Numerical Analysis*, vol. 7, Amsterdam: Elsevier.
- Cuyppers Y, Maurel A, and Petitjeans P (2003) *Physics Review Letters* 91: 194502.
- Dahmen W (1997) Wavelets and multiscale methods for operator equations. *Acta Numerica* 6: 55–228.
- Daubechies I (1992) *Ten Lectures on Wavelets*. SIAM.
- Daubechies I, Grossmann A, and Meyer Y (1986) *Journal of Mathematical Physics* 27: 1271.
- Farge M (1992) Wavelet transforms and their applications to turbulence. *Annual Reviews of Fluid Mechanics* 24: 395–457.
- Farge M and Rabreau G (1988) *Comptes Rendus Hebdomadaires des Seances de l'Academie des Sciences, Paris* 2: 307.
- Farge M, Kevlahan N, Perrier V, and Goirand E (1996) Wavelets and turbulence. *Proceedings of the IEEE* 84(4): 639–669.
- Farge M, Kevlahan N, Perrier V, and Schneider K (1999) Turbulence analysis, modelling and computing using wavelets. In: van den Berg JC (ed.) *Wavelets in Physics*, pp. 117–200. Cambridge: Cambridge University Press.
- Farge M and Schneider K (2002) Analysing and computing turbulent flows using wavelets. In: Lesieur M, Yaglom A, and David F (eds.) *New trends in turbulence*, Les Houches 2000, vol. 74, pp. 449–503. Springer.
- Farge M, Schneider K, Pellegrino G, Wray AA, and Rogallo RS (2003) *Physics of Fluids* 15(10): 2886.
- Grossmann A and Morlet J (1984) *SIAM J. Appl. Anal* 15: 723.
- Lemarié P-G and Meyer Y (1986) *Revista Matematica Iberoamericana* 2: 1.
- Mallat S (1989) *Transactions of the American Mathematical Society* 315: 69.
- Mallat S (1998) *A Wavelet Tour of Signal Processing*. Academic Press.
- Schneider K, Farge M, and Kevlahan N (2004) Spatial intermittency in two-dimensional turbulence. In: Tongring N and Penner RC (eds.) *Woods Hole Mathematics, Perspectives in Mathematics and Physics*, pp. 302–328. Word Scientific.
- <http://wavelets.ens.fr> – other papers about wavelets and turbulence can be downloaded from this site.

Wavelets: Applications

M Yamada, Kyoto University, Kyoto, Japan

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Wavelet analysis was first developed in the early 1980s in the field of seismic signal analysis in the form of an integral transform with a localized kernel function with continuous parameters of dilation and translation. When a seismic wave or its derivative has a singular point, the integral transform has a scaling property with respect to the dilation parameter; thus, this scaling behavior can be available to locate the singular point. In the mid-1980s, the orthonormal smooth wavelet was first constructed, and later the construction method was generalized and reformulated as multiresolution analysis (MRA). Since then, several kinds of wavelets have been proposed for various purposes, and the concept of wavelet has been extended to new types of basis functions. In this sense, the most important effect of wavelets may be that they have awakened deep interest in bases employed in data analysis and data processing. Wavelets are now widely used in various

fields of research; some of their applications are discussed in this article.

From the perspective of time–frequency analysis, the wavelet analysis may be regarded as a windowed Fourier analysis with a variable window width, narrower for higher frequency. The wavelets can therefore give information on the local frequency structure of an event; they have been applied to various kinds of one-dimensional (1D) or multi-dimensional signals, for example, to identify an event or to denoise or to sharpen the signal.

1D wavelets $\psi^{(a,b)}(x)$ are defined as

$$\psi^{(a,b)}(x) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{x-b}{a}\right)$$

where $a(\neq 0), b$ are real parameters and $\psi(x)$ is a spatially localized function called “analyzing wavelet” or “mother wavelet.” Wavelet analysis gives a decomposition of a function into a linear combination of those wavelets, where a perfect reconstruction requires the analyzing wavelet to satisfy some mathematical conditions.

For the continuous wavelet transform (CWT), where the parameters (a, b) are continuous, the

analyzing wavelet $\psi(x)L^2(\mathbf{R})$ has to satisfy the admissibility condition

$$C_\psi \equiv \int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty$$

where $\hat{\psi}(\omega)$ is the Fourier transform of $\psi(x)$:

$$\hat{\psi}(\omega) = \int_{-\infty}^{\infty} e^{-i\omega x} \psi(x) dx$$

The admissibility condition is known to be equivalent to the condition that $\psi(x)$ has no zero-frequency component, that is, $\hat{\psi}(0)=0$, under some mild condition for the decay rate at infinity. Then the CWT and its inverse transform of a data function $f(x) \in L^2(\mathbf{R})$ is defined as

$$T_\psi(a, b) = \frac{1}{\sqrt{C_\psi}} \int_{-\infty}^{\infty} \overline{\psi^{(a,b)}(x)} f(x) dx$$

$$f(x) = \frac{1}{\sqrt{C_\psi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} T_\psi(a, b) \psi^{(a,b)}(x) \frac{da db}{a^2}$$

In the case of the discrete wavelet transform (DWT), the parameters (a, b) are taken discrete; a typical choice is $a = 1/2^j$, $b = k/2^j$, where j and k are integers:

$$\psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k)$$

In order that the wavelets $\{\psi_{j,k}(x) | j, k \in \mathbf{Z}\}$ may constitute a complete orthonormal system in $L^2(\mathbf{R})$, the analyzing wavelet should satisfy more stringent conditions than the admissibility condition for the CWT, and is now constructed in the framework of MRA. A data function is then decomposed by the DWT as

$$f(x) = \sum_{j=-\infty}^{\infty} \alpha_{j,k} \psi_{j,k}(x), \quad \alpha_{j,k} = \int_{-\infty}^{\infty} \overline{\psi_{j,k}(x)} f(x) dx$$

Even when the discrete wavelets do not constitute a complete orthonormal system, they often form a wavelet frame if linear combinations of the wavelets are dense in $L^2(\mathbf{R})$ and if there are two constants A, B such that the inequality

$$A\|f\|^2 \leq \sum_{j,k} |\langle \psi_{j,k}, f \rangle|^2 \leq B\|f\|^2$$

holds for an arbitrary $f(x) \in L^2(\mathbf{R})$. For the wavelet frame $\{\psi_{j,k}\}$, there is a corresponding dual frame, $\{\tilde{\psi}_{j,k}\}$, which permits the following expansion of $f(x)$:

$$f(x) = \sum_{j,k} \langle \psi_{j,k}, f \rangle \tilde{\psi}_{j,k}(x) = \sum_{j,k} \langle \tilde{\psi}_{j,k}, f \rangle \psi_{j,k}(x)$$

The wavelet frame is also employed in several applications.

From the prospect of applications, the CWTs are better adapted for the analysis of data functions, including the detection of singularities and patterns, while the DWTs are adapted to the data processing, including signal compression or denoising.

Singularity Detection and Multifractal Analysis of Functions

Since its birth, the wavelet analysis has been applied for the detection of singularity of a data function. Let us define the Hölder exponent $h(x_0)$ at x_0 of a function $f(x)$ is defined here as the largest value of the exponent h such that there exists a polynomial $P_n(x)$ of degree n that satisfies for x in the neighborhood of x_0 :

$$|f(x) - P_n(x - x_0)| = O(|x - x_0|^h)$$

The data function is not differentiable if $h(x_0) < 1$, but if $h(x_0) > 1$ then it is differentiable and a singularity may arise in its higher derivatives. The wavelet transform is applied to find the Hölder exponent $h(x_0)$, because $T_\psi(a, b)$ has an asymptotic behavior $T_\psi(a, b) = O(a^{h(x_0+1/2)})(a \rightarrow 0)$ if the analyzing wavelet has $N(>h(x_0))$ vanishing moments, that is,

$$\int_{-\infty}^{\infty} x^m \psi(x) dx = 0, \quad m \in \mathbf{Z}, 0 \leq m < N$$

A commonly used analyzing wavelet for this purpose may be the N -time derivative of the Gaussian function $\psi(x) = d^N(e^{-x^2/2})/dx^N$. This method works well to examine a single or some finite number of singular points of the data function.

When the data function is a multifractal function with an infinite number of singular point of various strengths, the multifractal property of the data function is often characterized by the singularity spectrum $D(h)$ which denotes the Hausdorff dimension of the set of points where $h(x)=h$. The singularity spectrum is, however, difficult to obtain directly from the CWT, and the Legendre transformation is introduced to bypass the difficulty.

Fully developed 3D fluid turbulence may be a typical example of wavelet application to the singularity detection. The Kolmogorov similarity law of fluid turbulence for the longitudinal velocity increment $\Delta u(r) \equiv \mathbf{e} \cdot (\mathbf{u}(\mathbf{x} + r\mathbf{e}) - \mathbf{u}(\mathbf{x}))$, where $\mathbf{u}(\mathbf{x})$ is the velocity field and \mathbf{e} is a constant unit vector,

predicts a scaling property of the structure function; for r in the inertial subrange,

$$\langle (\Delta u(r))^p \rangle \sim r^{\zeta_p}, \quad \zeta_p = p/3$$

where $\langle \cdot \rangle$ denotes the statistical mean. In reality, however, the scaling exponent ζ_p measured in experiments shows a systematic deviation from $p/3$, which is considered to be a reflection of intermittency, namely the spatial nonuniformity or multifractal property of active vortical motions in turbulence. For simplicity, let us consider the velocity field on a linear section of the turbulence field. According to the multifractal formalism, the turbulence velocity field has singularities of various strengths described by the singularity spectrum $D(h)$, which is related to the scaling exponent ζ_p through the Legendre transform, $D(h) = \inf_p (ph - \zeta_p + 1)$. This relation is often used to determine $D(h)$ from the knowledge of ζ_p (structure function method). However, this method does not necessarily work well because, for example, it does not capture the singular points of the Hölder exponent larger than 1 and it is unstable for $h < 0$.

These difficulties are not restricted to the turbulence research, but arise commonly when the structure function is employed to determine the singularity spectrum. In these problems, the CWT $T_\psi(a, b)$ provides an alternative method. An ingenious technique is to take only the modulus maxima of $T_\psi(a, b)$ (for each of fixed a) to construct a partition function

$$Z(a, q) = \sum_{l \in L_{\max}} \left[\sup_{(a, b') \in l} |T_\psi(a, b')| \right]^q$$

where $q \in \mathbf{R}$, and L_{\max} denotes the set of all maxima lines, each of which is a continuous curve for small value of a , and there exists at least one maxima line toward a singular point of the Hölder exponent $h(x_0) < N$. In the limit of $a \rightarrow 0$, defining the exponent $\tau(q)$ as $Z(a, q) \sim a^{\tau(q)}$, one can obtain the singularity spectrum through the Legendre transform:

$$D(h) = \inf_q \left[q \left(h + \frac{1}{2} \right) - \tau(q) \right]$$

This method (wavelet-transform modulus-maxima (WTMM) method) is advantageous in that it works also for singularities of $h > 1$ and $h < 0$. Several simple examples of multifractal functions have been successfully analyzed by this method. For fluid turbulence, this method gives a singularity spectrum $D(h)$ which has a peak value of ~ 1 at $h \sim 1/3$, consistently with Kolmogorov similarity law, but

has a convex shape around $h = 1/3$ suggesting a multifractal property. For a fractal signal, we note that the WTMM method enlightens the hierarchical organization of the singularities, in the branching structure of the WT skeleton defined by the maxima lines arrangement in the (a, b) half-plane.

Though the above discussion also applies to the DWT, the detection of the Hölder exponent h in experimental situations is usually performed by the CWT, which has no restriction on possible values of a , while the DWT is often employed for theoretical discussions of singularity and multifractal structure of a function.

Multiscale Analysis

Wavelet transform expands a data function in the time–frequency or the position–wavenumber space, which has twice the dimension of the original signal, and makes it easier to perform a multiscale analysis and to identify events involved in the signal. In the wavelet transform, as stated above, the time resolution is higher at higher frequency, in contrast with the windowed Fourier transform where the time and the frequency resolutions are independent of frequency. Another advantage of wavelet is a wide variety of analyzing wavelet, which enables us to optimize the wavelet according to the purpose of data analysis. Both the CWT and the DWT are available for these time–frequency or position–wavenumber analysis. However, the CWT has properties quite different from those of familiar orthonormal bases of discrete wavelets.

Multidimensional CWT

The CWT can be formulated in an abstract way. We can regard $G = \{(a, b) | a \neq 0, b \in \mathbf{R}\}$ as an affine group on \mathbf{R} with the group operation of $(a, b)(a', b') = (aa', ab' + b)$ associated with the invariant measure $d\mu = da db/a^2$. The group G has its unitary representation in the Hilbert space $H = L^2(\mathbf{R})$:

$$(U(a, b)f)(x) = \frac{1}{\sqrt{|a|}} f\left(\frac{x - b}{a}\right)$$

and then we can consider the CWT can be constructed as a linear map W from $L^2(\mathbf{R})$ to $L^2(G; da db/a^2)$:

$$W : f(x) \mapsto T_\psi(a, b) = \frac{1}{\sqrt{C_\psi}} \langle U(a, b)\psi, f \rangle$$

where $\langle \cdot, \cdot \rangle$ is the inner product of $L^2(\mathbf{R})$ with the complex conjugate taken at the first element, and

$\psi(x)$ is a unit vector (analyzing wavelet) satisfying the abstract admissibility condition

$$C_\psi = \int_G |\langle U(a, b)\psi, \psi \rangle|^2 d\mu < \infty$$

This formulation is applicable also to a locally compact group G and its unitary and square integrable representation in a Hilbert space H . Note that even the canonical coherent states are included in this framework by taking the Weyl-Heisenberg group and $L^2(\mathbf{R})$ for G and H , respectively. This abstract formulation allows us to extend the CWT to higher-dimensional Euclidean spaces and other manifolds: for example, 2D sphere S^2 for geophysical application and 4D manifold of spacetime taking the Poincaré group into consideration.

In \mathbf{R}^n , the CWT of $f(x) \in L^2(\mathbf{R}^n)$ and its inverse transform are given by

$$T_\psi(a, r, b) = \frac{1}{\sqrt{C_\psi}} \int_{\mathbf{R}^n} \overline{\psi^{(a, r, b)}(x)} f(x) dx$$

$$f(x) = \frac{1}{\sqrt{C_\psi}} \int_G T(a, r, b) \psi^{(a, r, b)}(x) \frac{da dr db}{a^{n+1}}$$

where $r \in \text{SO}(n)$, $b \in \mathbf{R}^n$, dr is the normalized invariant measure of $G = \text{SO}(n)$, and the wavelets are defined as $\psi^{(a, r, b)}(x) = (1/a^{n/2})\psi(r^{-1}(x - b)/a)$, with the analyzing wavelet satisfying the admissibility condition

$$C_\psi = \int_{\mathbf{R}^n} \frac{|\hat{\psi}(\omega)|}{|\omega|^n} d\omega < \infty$$

Note that these wavelets are constructed not only by dilation and translation but also by rotation which therefore gives the possibility for directional pattern detection in a data function. In the case of 2D sphere S^2 , on the other hand, the dilation operation should be reinterpreted in such a way that at the North Pole, for example, it is the normal dilation in the tangent plane followed by lifting it to S^2 by the stereographic projection from the South Pole.

Generally, the abstract map W thus defined is injective and therefore reversal, but not surjective in contrast with the Fourier case. Actually in the case of 1D CWT, $T_\psi(a, b)$ is subject to an integral condition:

$$T_\psi(a, b) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{da db}{a^2} K(a, b; a', b') T_\psi(a', b')$$

$$K(a, b; a', b') = \int_{-\infty}^{\infty} \overline{\psi^{(a, b)}(x)} \psi^{(a', b')}(x) dx$$

which defines the range of the CWT, a subspace of $L^2(\mathbf{R})$. Therefore, if one wants to modify $T_\psi(a, b)$ by, for example, assigning its value as zero in some parameter region just as in a filter process, care should be taken for the resultant $T_\psi(a, b)$ to be in the image of the CWT. The reason may be understood intuitively by noticing that the wavelets $\psi^{(a, b)}(x)$ are linearly dependent on each other. The expression of a data function by a linear combination of the wavelets is therefore not unique, and thus is redundant. The CWT gives only $T_\psi(a, b)$ of the least norm in $L^2(\mathbf{R}^2; da db/a^2)$. In physical interpretations of the CWT, however, this nonuniqueness is often ignored.

Pattern Detection

Edge detection The edges of an object are often the most important components for pattern detection. The edge may be considered to consist of points of sharp transition of image intensity. At the edge, the modulus of the gradient of the image $f(x, y)$ is expected to take a local maximum in the 1D direction perpendicular to the edge. Therefore, the local maxima of $|\nabla f(x, y)|$ may be the indicator of the edge. However, the image textures can also give similar sharp transitions of $f(x, y)$, and one should take into account the scale dependence which distinguishes between edges and textures. One of the practically possible ways for this purpose is to use dyadic wavelets $\psi_j^m(x, y) = 2^j \psi^m(2^j x, 2^j y)$ which are generated from the two wavelets $(\psi^1, \psi^2) = (-\partial\theta/\partial x, -\partial\theta/\partial y)$, where θ is a localized function (multiscale edge detection method). The dyadic wavelet transform of the image $f(x, y)$

$$T_j^m(b_1, b_2) = \langle f(x, y), \psi_j^m(x - b_1, y - b_2) \rangle, \quad m = 1, 2$$

defines the multiscale edges as a set of points $b = (b_1, b_2)$ where the modulus of the wavelet transform, $|(T_j^1, T_j^2)|$, takes a locally maximum value (WTMM) in a 1D neighborhood of b in the direction of $(T_j^1(b), T_j^2(b))$. Scale dependence of the magnitude of the modulus maxima is related to the Hölder exponent of $f(x, y)$ similarly to 1D case, and thus gives information to distinguish between the edges and the textures.

Inversely, the information of WTMM $b_{j,p} = \{(b_{1,j,p}, b_{2,j,p})\}$ of multiscale edges can be made use of for an approximate reconstruction of the original image, although the perfect reconstruction cannot be expected because of the noncompleteness of the modulus maxima wavelets. Assuming that $\{\psi_{j,p}^1, \psi_{j,p}^2\} = \{\psi_j^1(x - b_{j,p}), \psi_j^2(x - b_{j,p})\}$ constitutes a frame of the linear closed space generated by

$\{\psi_{i,p}^1, \psi_{i,p}^2\}$, an approximate image \hat{f} is obtained by inverting the relation

$$L\hat{f} \equiv \sum_m \sum_{j,p} \langle \hat{f}, \psi_{j,p}^m \rangle \psi_{j,p}^m = \sum_m \sum_{j,p} T_j^m(b_{j,p}) \psi_{j,p}^m$$

using, for example, a conjugate gradient algorithm, where a fast calculation is possible with a filter bank algorithm for the dyadic wavelet ("algorithm à trous"). This algorithm gives only the solution of minimum norm among all possible solutions, but it is often satisfactory for practical purposes and thus is applicable also to data compression.

Directional detection For oriented features such as segments or edges in images to be detected, a directionally selective wavelet for the CWT is desired. A useful wavelet for this purpose is one that has the effective support of its Fourier transform in a convex cone with apex at the origin in wave number space. A typical example of the directional wavelet may be the 2D Morlet wavelet:

$$\psi(\mathbf{x}) = \exp(i\mathbf{k}_0 \cdot \mathbf{x}) \exp(-|A\mathbf{x}|^2)$$

where \mathbf{k}_0 is the center of the support in Fourier space, and A is a 2×2 matrix $\text{diag}[\epsilon^{1/2}, 1]$ ($\epsilon \leq 1$), where the admissibility condition for the CWT is approximately satisfied for $|\mathbf{k}_0| \geq 5$. Another example is the Cauchy wavelet which has the support strictly in a convex cone in wave number space.

These wavelets have the directional selectivity with preference to a slender object in a specific direction. One of their applications is the analysis of the velocity field of fluid motion from an experimental data, where many tiny plastic balls distributed in fluid give a lot of line segments in a picture taken with a short exposure. The directional wavelet analysis of the picture classifies the line segments according to their directions, indicating the directions of fluid velocity. Another example may be a wave-field analysis where many waves in different directions are superimposed; the directional wavelets allow one to decompose the wave field into the component waves. Directional wavelets have also been applied successfully to detect symmetry of objects such as crystals or quasicrystals.

Denoising and separation of signals The wavelet frame as well as the CWT give a redundant representation of a data function. If, instead of the original data, the redundant expression is transmitted, the redundancy is used to reduce the noise included in the received data because the redundancy requires the data to belong to a subspace, and the projection of the received data to the subspace

reduces the noise component orthogonal to it. More specifically, the wavelet frame gives a representation of a data function as $f(t) = \sum_{j,k} \alpha_{j,k} \psi_{j,k}$, where the expansion coefficients $\alpha_{j,k} = \langle \psi_{j,k}, f(x) \rangle$ satisfy the defining equation of the subspace

$$\alpha_{j',k'} = \sum \alpha_{j,k} \langle \psi_{j',k'}, \psi_{j,k} \rangle$$

If the frame coefficients are transmitted, the projection operator P , which is defined on the right-hand side of the above equation, reduces the noise in the received coefficients $\alpha_{j,k}$ contaminated during the transmission.

However, this method is not applicable if the transmitted signal is not redundant. Then some *a priori* criterion is necessary to discriminate between signal and noise. Various criteria have been proposed in different fields. If the signal and the noise, or plural signals have different power-law forms of spectra, then their discrimination may be possible by the DWT at higher-frequency region where the difference in the magnitude of the coefficients is significant. In this approach, the wavelets of Meyer type, that is, an orthogonal wavelet with a compact support in Fourier space, may be preferable because the wavelets of different scales are separated, at least to some extent, in Fourier space.

In fluid dynamics, the vorticity field of 2D turbulence is found to be decomposed into coherent and incoherent vorticity fields, according as the CWT is larger than a threshold value or not, respectively. These two fields give different Fourier spectra of the velocity field (k^{-5} for coherent part while k^{-3} for incoherent part), showing that the coherent structures are responsible for the deviation from k^{-3} predicted by the classical enstrophy cascade theory. In an astronomical application, on the other hand, the data processing is performed by a more sophisticated method taking into account interscale relation in the wavelet transform, because an astronomical image contains various kinds of objects, including stars, double-stars, galaxies, nebulae, and clusters. In a medical image however contrast analysis is indispensable for diagnostic imaging to get a clear detailed picture of organic structure. A scale-dependent local contrast is defined as the ratio of the CWT to that given by an analyzing wavelet with a larger support. A multiplicative scheme to improve the contrast is constructed by using the local contrast.

Signal Compression

Signal compression is quite an important technology in digital communication. Speech, audio, image, and digital video are all important fields of signal

compression, and plenty of compression methods have been put to practical use, but we mention here only a few.

The MRA for orthogonal wavelets gives a successive procedure to decompose a subspace of $L^2(\mathbf{R})$ into a direct sum of two subspaces corresponding to higher- and lower-frequency parts; only the latter of which is decomposed again into its higher- and lower-frequency parts. Algebraically, this procedure was already known before the discovery of MRA in filter theory in electrical engineering, where a discretely sampled signal is convoluted with a filter series to give, for example, a high-pass-filtered or low-pass-filtered series. An appropriate designed pair of a high-pass and a low-pass filters followed by the downsampling yields two new series corresponding to the higher- and lower-frequency parts, respectively, which are then reversible by another two reconstruction filters with the upsampling. These four filters which are often employed in a widely used technique of “sub-band coding” then constitute a perfect reconstruction filter bank. Under some conditions, successive applications of this decomposition process to the series of lower-frequency parts, which is equivalent to the nesting structure of MRA, have been used for data compression (quadrature mirror filter). A famous example is a data compression system of FBI for finger prints, consisting of wavelet coding with scalar quantization.

In MRA, however, it is only the lower-frequency parts that are successively decomposed. If both the lower- and the higher-frequency parts are repeatedly decomposed by the decomposition filters, then the successive convolution processes correspond to a decomposition of data function by a set of wavelet-like functions, called “wavelet packet,” where there are choices whether to decompose the higher- and/or the lower-frequency parts. The best wavelet packet, in the sense of the entropy, for example, within a specified number of decompositions, often provides with a powerful tool for data compression in several areas, including speech analysis and image analysis. We also note that from the viewpoint of the best basis which minimizes the statistical mean square error of the thresholded coefficients, an orthonormal wavelet basis gives a good concentration of the energy if the original signal is a piecewise smooth function superimposed by a white noise, which is thus efficiently removed by thresholding the coefficients. The efficiency of a wavelet expansion of a signal is sometimes evaluated with the entropy of “probability” defined as $|\alpha_{j,k}|^2/||f||^2$. A better wavelet can be selected by reducing the entropy, practically from among some set of wavelets, and its restricted expansion coefficients

give a compressed signal. One of the systematic methods to generate such a suitable basis is also to employ the wavelet packets.

Numerical Calculation

Application of wavelet transform, especially of the DWT, to numerical solver for a differential equation (DE) has long been studied. At the first sight, the wavelets appear to give a good DE solver because the wavelet expansion is generally quite efficient compared to Fourier series due to its spatial localization. But its implementation to an efficient computer code is not so straightforward; research is still continuing for concrete problems. Application of the CWT to spectral method for partial differential equation (PDE) has been studied extensively. There is no wavelet which diagonalizes the differential operator $\partial/\partial x$; therefore, an efficient numerical method is necessary for derivatives of wavelets. Products of wavelets also yield another numerical problem. MRA brings about mesh points which are adaptive to some extent, but finite element method still gives more flexible mesh points.

For some scaling-invariant differential or integral operators, including $\partial^2/\partial x^2$, Abel transformations, and Reisz potential, adaptive biorthogonal wavelets can be provided with block-diagonal Galerkin representations, which has been applied to data processing. Generally, simultaneous localization of wavelets, both in space and in scale, leads to a sparse Galerkin representation for many pseudodifferential operators and their inverses. A thresholding technique with DWT has been introduced to coherent vortex simulation of the 2D Navier–Stokes equations, to reduce the relevant wavelet coefficients. Another promising application of wavelet occurs as a preprocessor for an iterative Poisson solver, where a wavelet-based preconditioning leads to a matrix with a bounded condition number.

Other Wavelets and Generalizations

Several new types of wavelets have been proposed: “coiflet” whose scaling function has vanishing moments giving expansion coefficients approximately equal to values of the data functions, and “symlet” which is an orthonormal wavelet with a nearly symmetric profile. Multiwavelets are wavelets which give a complete orthonormal system in L^2 space. In 2D or multidimensional applications of the DWT, separable orthonormal wavelets consisting of tensor products of 1D orthonormal wavelets are frequently used, while nonseparable orthonormal wavelets are also available. Another generalization

of wavelets is the Malvar basis which is also a generalization of local Fourier basis, and gives a perfect reconstruction. A new direction of wavelet is the second-generation wavelets which are constructed by lifting scheme and free from the regular dyadic procedure, and thus applicable to compact regions as S^2 and a finite interval.

See also: Fractal Dimensions in Dynamics; Image Processing: Mathematics; Intermittency in Turbulence; Wavelets: Application to Turbulence; Wavelets: Mathematical Theory.

Wavelets: Mathematical Theory

K Schneider, Université de Provence, Marseille, France

M Farge, Ecole Normale Supérieure, Paris, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The wavelet transform unfolds functions into time (or space) and scale, and possibly directions. The continuous wavelet transform has been discovered by Alex Grossmann and Jean Morlet who published the first paper on wavelets in 1984. This mathematical technique, based on group theory and square-integrable representations, allows us to decompose a signal, or a field, into both space and scale, and possibly directions. The orthogonal wavelet transform has been discovered by Lemarié and Meyer (1986). Then, Daubechies (1988) found orthogonal bases made of compactly supported wavelets, and Mallat (1989) designed the fast wavelet transform (FWT) algorithm. Further developments were done in 1991 by Raffy Coifman, Yves Meyer, and Victor Wickerhauser who introduced wavelet packets and applied them to data compression. The development of wavelets has been interdisciplinary, with contributions coming from very different fields such as engineering (sub-band coding, quadrature mirror filters, time–frequency analysis), theoretical physics (coherent states of affine groups in quantum mechanics), and mathematics (Calderon–Zygmund operators, characterization of function spaces, harmonic analysis). Many reference textbooks are available, some of them we recommend are listed in the “Further reading” section. Meanwhile, a large spectrum of applications has grown and is still developing, ranging from signal analysis and image processing via numerical analysis and turbulence modeling to data compression.

Further Reading

- Benedetto JJ and Frazier W (eds.) (1994) *Wavelets: Mathematics and Applications*. Boca Raton, FL: CRC Press.
- van den Berg JC (ed.) (1999) *Wavelets in Physics*. Cambridge: Cambridge University Press.
- Daubechies I (1992) *Ten Lectures on Wavelets*, SIAM, CBMS61, Philadelphia.
- Mallat S (1998) *A Wavelet Tour of Signal Processing*. San Diego: Academic Press.
- Strang G and Nguyen T (1997) *Wavelet and Filter Banks*. Wellesley: Wellesley-Cambridge Press.

In this article, we will first define the continuous wavelet transform and then the orthogonal wavelet transform based on a multiresolution analysis. Properties of both transforms will be discussed and illustrated by examples. For a general introduction to wavelets, see *Wavelets: Applications*.

Continuous Wavelet Transform

Let us consider the Hilbert space of square-integrable functions $L^2(\mathbb{R}) = \{f : \|f\|_2 < \infty\}$, equipped with the scalar product $\langle f, g \rangle = \int_{\mathbb{R}} f(x)g^*(x) dx$ (* denotes the complex conjugate in the case of complex-valued functions) and where the norm is defined by $\|f\|_2 = \langle f, f \rangle^{1/2}$.

Analyzing Wavelet

The starting point for the wavelet transform is to choose a real- or complex-valued function $\psi \in L^2(\mathbb{R})$, called the “mother wavelet,” which fulfills the admissibility condition,

$$C_\psi = \int_0^\infty \left| \widehat{\psi}(k) \right|^2 \frac{dk}{|k|} < \infty \quad [1]$$

where

$$\widehat{\psi}(k) = \int_{-\infty}^\infty \psi(x) e^{-i2\pi kx} dx \quad [2]$$

denotes the Fourier transform, with $i = \sqrt{-1}$ and k the wave number. If ψ is integrable, that is, $\psi \in L^1(\mathbb{R})$, this implies that ψ has zero mean,

$$\int_{-\infty}^\infty \psi(x) dx = 0 \quad \text{or} \quad \widehat{\psi}(0) = 0 \quad [3]$$

In practice, however, one also requires the wavelet ψ to be well localized in both physical and Fourier

spaces, the latter implying smoothness, and to have M vanishing moments,

$$\int_{-\infty}^{\infty} x^m \psi(x) dx = 0 \quad \text{for } m = 0, M-1 \quad [4]$$

that is, monomials up to degree $M-1$ are exactly reproduced. In Fourier space, this property is equivalent to

$$\frac{d^m}{dk^m} \widehat{\psi}(k) |_{k=0} = 0 \quad \text{for } m = 0, M-1 \quad [5]$$

therefore, the Fourier transform of ψ decays smoothly at $k=0$.

Analysis

From the mother wavelet ψ , we generate a family of continuously translated and dilated wavelets,

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}} \psi\left(\frac{x-b}{a}\right) \quad \text{for } a > 0 \text{ and } b \in \mathbb{R} \quad [6]$$

where a denotes the dilation parameter, corresponding to the width of the wavelet support, and b the translation parameter, corresponding to the position of the wavelet. The wavelets are normalized in energy norm, that is, $\|\psi_{a,b}\|_2 = 1$.

In Fourier space, eqn [6] reads

$$\widehat{\psi}_{a,b}(k) = \sqrt{a} \widehat{\psi}(ak) e^{-i2\pi kb} \quad [7]$$

where the contraction with $1/a$ in [6] is reflected in a dilation by a [7] and the translation by b implies a rotation in the complex plane.

The continuous wavelet transform of a function f is then defined as the convolution of f with the wavelet family $\psi_{a,b}$:

$$\widetilde{f}(a,b) = \int_{-\infty}^{\infty} f(x) \psi_{a,b}^*(x) dx \quad [8]$$

where $\psi_{a,b}^*$ denotes, in the case of complex-valued wavelets, the complex conjugate.

Using Parseval's identity, we get

$$\widetilde{f}(a,b) = \int_{-\infty}^{\infty} \widehat{f}(k) \widehat{\psi}_{a,b}^*(k) dk \quad [9]$$

and the wavelet transform could be interpreted as a frequency decomposition using bandpass filters $\widehat{\psi}_{a,b}$ centered at frequencies $k = k_\psi/a$. The wave number k_ψ denotes the barycenter of the wavelet support in Fourier space

$$k_\psi = \frac{\int_0^\infty k |\widehat{\psi}(k)| dk}{\int_0^\infty |\widehat{\psi}(k)| dk} \quad [10]$$

Note that these filters have a variable width $\Delta k/k$; therefore, when the wave number increases, the bandwidth becomes wider.

Synthesis

The admissibility condition [1] implies the existence of a finite energy reproducing kernel, which is a necessary condition for being able to reconstruct the function f from its wavelet coefficients \widetilde{f} . One then recovers

$$f(x) = \frac{1}{C_\psi} \int_0^\infty \int_{-\infty}^\infty \widetilde{f}(a,b) \psi_{a,b}(x) \frac{da db}{a^2} \quad [11]$$

which is the inverse wavelet transform.

The wavelet transform is an isometry and one has Parseval's identity. Therefore, the wavelet transform conserves the inner product and we obtain

$$\begin{aligned} \langle f, g \rangle &= \int_{-\infty}^\infty f(x) g^*(x) dx \\ &= \frac{1}{C_\psi} \int_0^\infty \int_{-\infty}^\infty \widetilde{f}(a,b) \widetilde{g}^*(a,b) \frac{da db}{a^2} \end{aligned} \quad [12]$$

As a consequence, the total energy E of a signal can be calculated either in physical space or in wavelet space, such as

$$\begin{aligned} E &= \int_{-\infty}^\infty |f(x)|^2 dx \\ &= \frac{1}{C_\psi} \int_0^\infty \int_{-\infty}^\infty |\widetilde{f}(a,b)|^2 \frac{da db}{a^2} \end{aligned} \quad [13]$$

This formula is also the starting point for the definition of wavelet spectra and scalogram (see Wavelets: Application to Turbulence).

Examples

In the following, we apply the continuous wavelet transform to different academic signals using the Morlet wavelet. The Morlet wavelet is complex valued, and consists of a modulated Gaussian with width k_0/π :

$$\psi(x) = (e^{2i\pi x} - e^{-k_0^2/2}) e^{-2\pi^2 x^2/k_0^2} \quad [14]$$

The envelope factor k_0 controls the number of oscillations in the wave packet; typically, $k_0=5$ is used. The correction factor $e^{-k_0^2/2}$, to ensure its vanishing mean, is very small and often neglected. The Fourier transform is

$$\widehat{\psi}(k) = \frac{k_0}{2\sqrt{\pi}} e^{-(k_0^2/2)(1+k^2)} (e^{-k_0^2 k} - 1) \quad [15]$$

Figure 1 shows wavelet analyses of a cosine, two sines, a Dirac, and a characteristic function. Below the four signals we plot the modulus and the phase of the corresponding wavelet coefficients.

Higher Dimensions

The continuous wavelet transform can be extended to higher dimensions in $L^2(\mathbb{R}^n)$ in different ways. Either we define spherically symmetric wavelets by setting $\psi(x) = \psi^{1d}(|x|)$ for $x \in \mathbb{R}^n$ or we introduce in addition to dilations $a \in \mathbb{R}^+$ and translations $b \in \mathbb{R}^n$ also rotations to define wavelets with a directional sensitivity. In the two-dimensional case, we obtain for example,

$$\psi_{a,b,\theta}(x) = \frac{1}{a} \psi\left(R_\theta^{-1}\left(\frac{x-b}{a}\right)\right) \tag{16}$$

where $a \in \mathbb{R}^+, b \in \mathbb{R}^2$, and where R_θ is the rotation matrix

$$\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \tag{17}$$

The analysis formula [8] then becomes

$$\tilde{f}(a,b,\theta) = \int_{\mathbb{R}^2} f(x) \psi_{a,b,\theta}^*(x) \, dx \tag{18}$$

and for the corresponding inverse wavelet transform [11] we obtain

$$f(x) = \frac{1}{C_\psi} \int_0^\infty \int_{\mathbb{R}^2} \int_0^{2\pi} \tilde{f}(a,b,\theta) \psi_{a,b,\theta}(x) \frac{da db d\theta}{a^3} \tag{19}$$

Similar constructions can be made in dimensions larger than 2 using $n - 1$ angles of rotation.

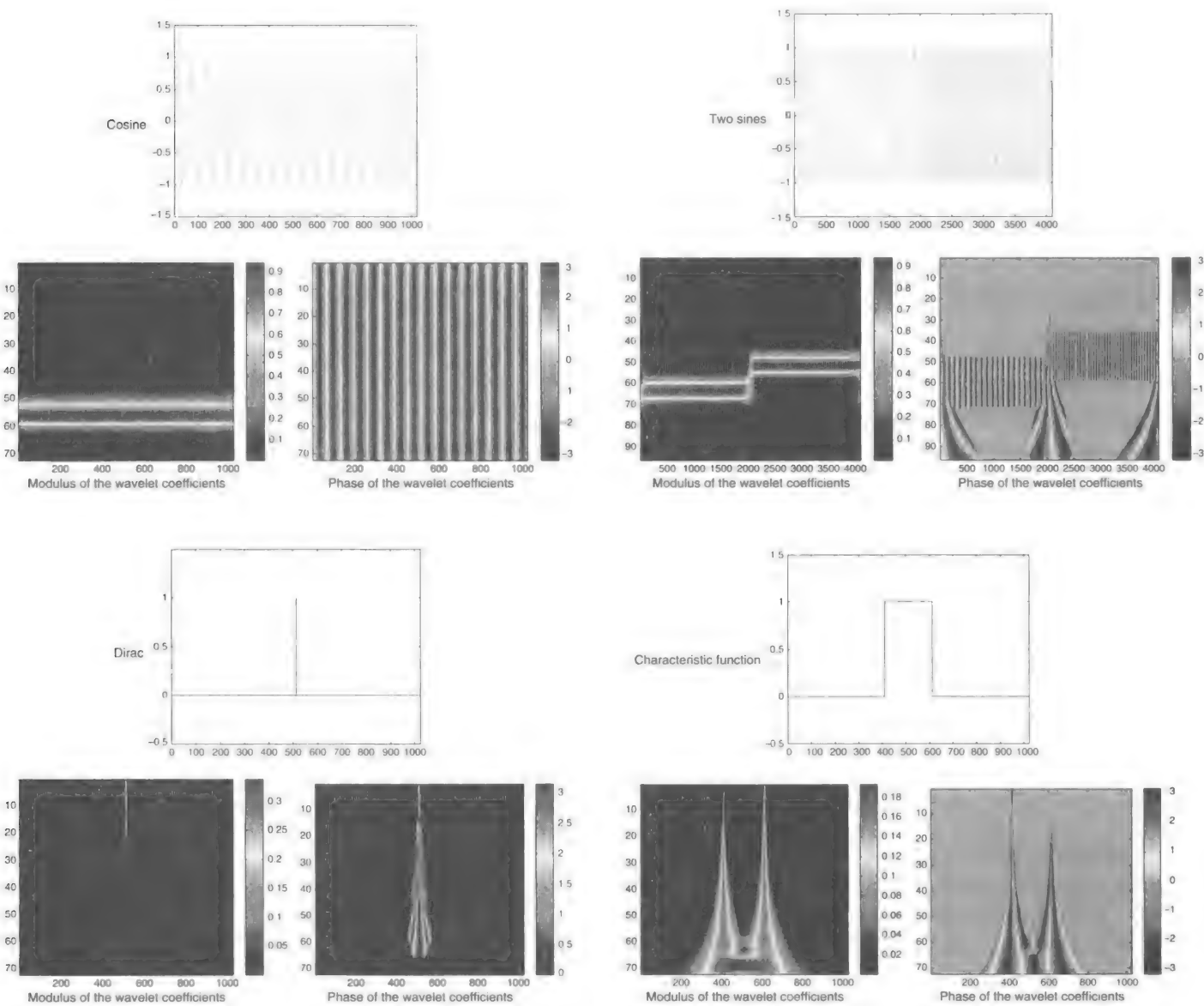


Figure 1 Examples of a one-dimensional continuous wavelet analysis using the complex-valued Morlet wavelet. Each subfigure shows on the top the function to be analyzed and below (left) the modulus of its wavelet coefficients and below (right) the phase of its wavelet coefficients.

Discrete Wavelets

Frames

It is possible to obtain a discrete set of quasiorthogonal wavelets by sampling the scale and position axes a, b . For the scale a we use a logarithmic discretization: a is replaced by $a_j = a_0^{-j}$, where a_0 is the sampling rate of the $\log a$ axis ($a_0 = \Delta(\log a)$) and where $j \in \mathbb{Z}$ is the scale index. The position b is discretized linearly: b is replaced by $x_{ji} = ib_0a_0^{-j}$, where b_0 is the sampling rate of the position axis at the largest scale and where $i \in \mathbb{Z}$ is the position index. Note that the sampling rate of the position varies with scale, that is, for finer scales (increasing j and hence decreasing a_j), the sampling rate increases. Accordingly, we obtain the discrete wavelets (cf. Figure 2)

$$\psi_{ji}(x') = a_j^{-1/2} \psi\left(\frac{x' - x_{ji}}{a_j}\right) \tag{20}$$

and the corresponding discrete decomposition formula is

$$\tilde{f}_{ji} = \langle \psi_{ji}, f \rangle = \int_{-\infty}^{\infty} f(x') \psi_{ji}^*(x') \, dx' \tag{21}$$

Furthermore, the wavelet coefficients satisfy the following estimate:

$$A \|f\|_2^2 \leq \sum_{j,i} |\tilde{f}_{ji}|^2 \leq B \|f\|_2^2 \tag{22}$$

with frame bounds $B \geq A > 0$. In the case $A = B$ we have a tight frame.

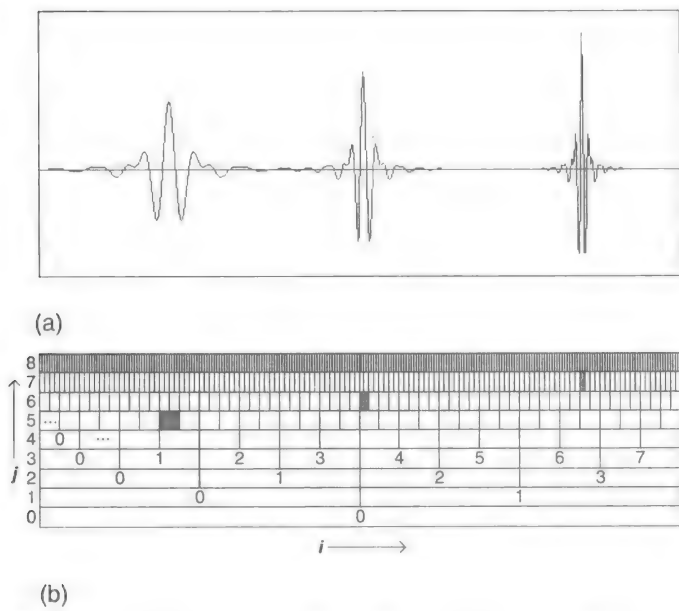


Figure 2 Orthogonal quintic spline wavelets $\psi_{j,i}(x) = 2^{j/2} \psi(2^j x - i)$ at different scales and positions: (a) $\psi_{5,6}(x)$, $\psi_{6,32}(x)$, $\psi_{7,108}(x)$, and (b) corresponding wavelet coefficients.

The discrete reconstruction formula is

$$f(x) = C \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} \tilde{f}_{ji} \psi_{ji}(x) + R(x) \tag{23}$$

where C is a constant and $R(x)$ is a residual, both depending on the choice of the wavelet and the sampling of the scale and position axes. For the particular choice $a_0 = 2$ (which corresponds to a scale sampling by octaves) and $b_0 = 1$, we have the dyadic sampling, for which there exist special wavelets ψ_{ji} that form an orthonormal basis of $L^2(\mathbb{R})$, that is, such that

$$\langle \psi_{ji}, \psi_{j'i'} \rangle = \delta_{jj'} \delta_{ii'} \tag{24}$$

where δ denotes the Kronecker symbol. This means that the wavelets ψ_{ji} are orthogonal with respect to their translates by discrete steps $2^{-j}i$ and their dilates by discrete steps 2^{-j} corresponding to octaves. In this case, the reconstruction formula is exact with $C = 1$ and $R = 0$. Note that the discrete wavelet transform has lost the invariance by translation and dilation of the continuous one.

Orthogonal Wavelets and Multiresolution Analysis

The construction of orthogonal wavelet bases and the associated fast numerical algorithm is based on the mathematical concept of multiresolution analysis (MRA). The underlying idea is to consider approximations f_j of the function f at different scales j . The amount of information needed to go from a coarse approximation f_j to a finer resolution approximation f_{j+1} is then described using orthogonal wavelets. The orthogonal wavelet analysis can thus be interpreted as decomposing the function into approximations of the function at coarser and coarser scales (i.e., for decreasing j), where the differences between the approximations are encoded using wavelets.

The definition of the MRA was introduced by Stéphane Mallat in 1988 (Mallat 1989). This technique constitutes a mathematical framework of orthogonal wavelets and the related FWT.

A one-dimensional orthogonal MRA of $L^2(\mathbb{R})$ is defined as a sequence of successive approximation spaces $V_j, j \in \mathbb{Z}$, which are closed imbedded subspaces of $L^2(\mathbb{R})$. They verify the following conditions:

$$V_j \subset V_{j+1} \quad \forall j \in \mathbb{Z} \tag{25}$$

$$\overline{\bigcup_{j \in \mathbb{Z}} V_j} = L^2(\mathbb{R}) \tag{26}$$

$$\bigcap_{j \in \mathbb{Z}} V_j = \{0\} \tag{27}$$

$$f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1} \tag{28}$$

A scaling function $\phi(x)$ is required to exist. Its translates generate a basis in each V_j , that is,

$$V_j V_j = \overline{\text{span}}\{\phi_{ji}\}_{i \in \mathbb{Z}} \quad [29]$$

where

$$\phi_{ji}(x) = 2^{j/2} \phi(2^j x - i), \quad j, i \in \mathbb{Z} \quad [30]$$

At a given scale j , this basis is orthonormal with respect to its translates by steps $i/2^j$ but not to its dilates,

$$\langle \phi_{ji}, \phi_{jk} \rangle = \delta_{ik} \quad [31]$$

The nestedness of the approximation spaces [28] generated by the scaling function ϕ implies that it satisfies a refinement equation:

$$\phi_{j-1,i}(x) = \sum_{n=-\infty}^{\infty} h_{n-2i} \phi_{jn}(x) \quad [32]$$

with the filter coefficients $h_n = \langle \phi_{jn}, \phi_{j-1,0} \rangle$, which determine the scaling function completely. In general, only the filter coefficients h_n are known and no analytical expression of ϕ is given. Equation [32] implies that the approximation of a function at coarser scale can be described by linear combinations of the same function at finer scales.

The orthogonal projection of a function $f \in L^2(\mathbb{R})$ on V_j is defined as

$$P_{V_j} : f \rightarrow P_{V_j} f = f_j \quad [33]$$

with

$$f_j(x) = \sum_{k \in \mathbb{Z}} \langle f, \phi_{jk} \rangle \phi_{jk}(x) \quad [34]$$

This coarse graining at a given scale J is done by filtering the function with the scaling function ϕ . As a filter, the scaling function ϕ does not have vanishing mean but is normalized so that $\int_{-\infty}^{\infty} \phi(x) dx = 1$.

As V_{j-1} is included in V_j , we can define its orthogonal complement space in V_j :

$$V_j = V_{j-1} \oplus W_{j-1} \quad [35]$$

Correspondingly, the approximation of the function f at scale 2^{-j} , belonging to V_j , can be decomposed as a sum of orthogonal projections on V_{j-1} and W_{j-1} , such that

$$P_{V_j} f = P_{V_{j-1}} f + P_{W_{j-1}} f \quad [36]$$

Based on the scaling function ϕ , one can construct a function ψ , the so-called mother wavelet, given by the relation

$$\psi_{ji}(x) = \sum_{n \in \mathbb{Z}} g_{n-2i} \phi_{jn}(x) \quad [37]$$

with $g_n = \langle \phi_{jn}, \psi_{j-1,0} \rangle$, and where $\psi_{ji}(x) = 2^{j/2} \psi(2^j x - i)$, $j, i \in \mathbb{Z}$ (cf. Figure 2). The filter coefficients g_n can be computed from the filter coefficients h_n using the relation

$$g_n = (-1)^{1-n} h_{1-n} \quad [38]$$

The translates and dilates of the wavelet ψ constitute orthonormal bases of the spaces W_j ,

$$W_j = \overline{\text{span}}\{\psi_{ji}\}_{i \in \mathbb{Z}} \quad [39]$$

As in the continuous case, the wavelets have vanishing mean, and also possibly vanishing higher-order moments; therefore,

$$\int_{-\infty}^{\infty} x^m \psi(x) dx = 0 \quad \text{for } m = 0, \dots, M-1 \quad [40]$$

Let us now consider approximations of a function $f \in L^2(\mathbb{R})$ at two different scales j :

- at scale j

$$f_j(x) = \sum_{i=-\infty}^{\infty} \bar{f}_{ji} \phi_{ji}(x) \quad [41]$$

- at scale $j-1$

$$f_{j-1}(x) = \sum_{i=-\infty}^{\infty} \bar{f}_{j-1,i} \phi_{j-1,i}(x) \quad [42]$$

with the scaling coefficients

$$\bar{f}_{ji} = \langle f, \phi_{ji} \rangle \quad [43]$$

which correspond to local averages of the function f at position $i2^{-j}$ and at scale 2^{-j} .

The difference between the two approximations is encoded by the wavelets

$$f_j(x) - f_{j-1}(x) = \sum_{i=-\infty}^{\infty} \tilde{f}_{j-1,i} \psi_{j-1,i}(x) \quad [44]$$

with the wavelet coefficients

$$\tilde{f}_{ji} = \langle f, \psi_{ji} \rangle \quad [45]$$

which correspond to local differences of the function at position $(2i+1)2^{-(j+1)}$ between approximations at scales 2^{-j} and $2^{-(j+1)}$.

Iterating the two-scale decomposition [44], any function $f \in L^2(\mathbb{R})$ can be expressed as a sum of a coarse-scale approximation at a reference scale j_0 that we set to 0 here, and their successive

differences. These details are needed to go from one scale j to the next finer scale $j+1$ for $j=0, \dots, J-1$,

$$f(x) = \sum_{i=-\infty}^{\infty} \bar{f}_{0,i} \phi_{0,i}(x) + \sum_{j=0}^{\infty} \sum_{i=-\infty}^{\infty} \tilde{f}_{ji} \psi_{ji}(x) \quad [46]$$

For numerical applications, the sums in eqn [46] have to be truncated in both scale j and position i . The truncation in scale corresponds to a limitation of f to a given finest scale J , which is in practice imposed by the available sampling rate. Due to the finite length of the available data, the sum over i also becomes finite. The decomposition [46] is orthogonal, as, by construction,

$$\langle \psi_{ji}, \psi_{j'i'} \rangle = \delta_{jj'} \delta_{ii'} \quad [47]$$

$$\langle \psi_{ji}, \phi_{j'i'} \rangle = 0 \quad \text{for } j \geq j' \quad [48]$$

in addition to [31].

Fast Wavelet Transform

Starting with a function $f \in L^2(\mathbb{R})$ given at the finest resolution 2^{-J} (i.e., we know $f_J \in V_J$ and hence the coefficients \bar{f}_{ji} for $i \in \mathbb{Z}$), the FWT computes its wavelet coefficients \tilde{f}_{ji} by decomposing successively each approximation f_j into a coarser scale approximation f_{j-1} , plus the corresponding details which are encoded by the wavelet coefficients. The algorithm uses a cascade of discrete convolutions with the low pass filter h_n and the bandpass filter g_n , followed by downsampling, in which only one coefficient out of two is retained. The direct wavelet transform algorithm is

- initialization

$$\text{given } f \in L^2(\mathbb{R}) \text{ and } \bar{f}_{ji} = f\left(\frac{i}{2^j}\right) \text{ for } i \in \mathbb{Z}$$

- decomposition

for $j=J$ to 1, step -1, do

$$\bar{f}_{j-1,i} = \sum_{n \in \mathbb{Z}} h_{n-2i} \bar{f}_{jn} \quad [49]$$

$$\tilde{f}_{j-1,i} = \sum_{n \in \mathbb{Z}} g_{n-2i} \bar{f}_{jn} \quad [50]$$

The inverse wavelet transform is based on successive reconstructions of fine-scale approximations f_j from coarser scale approximations f_{j-1} , plus the differences between approximations at scale $j-1$ and the finer scale j which are encoded by $\tilde{f}_{j-1,i}$. The algorithm uses a cascade of discrete convolutions with the filters h_n and g_n , preceded by

upsampling which adds zeros in between two successive coefficients.

- reconstruction

for $j=1$ to J , step 1, do

$$\bar{f}_{ji} = \sum_{n=-\infty}^{\infty} h_{i-2n} \bar{f}_{j-1,n} + \sum_{n=-\infty}^{\infty} g_{i-2n} \tilde{f}_{j,n} \quad [51]$$

The FWT has been introduced by Stéphane Mallat in 1989. If the scaling functions (and wavelets) are compactly supported, the filters h_n and g_n have only a finite number of nonvanishing coefficients. In this case, the numerical complexity of the FWT is $\mathcal{O}(N)$ where N denotes the number of samples.

Choice of Wavelets

Orthogonal wavelets are typically defined by their filter coefficients h_n , since in general no analytic expression for ψ is available. In the following, we give the filter coefficients of h_n for some typical orthogonal wavelets. The filter coefficients of g_n can be obtained using the quadrature relation between the two filters [38].

- Haar D1 (one vanishing moment):

$$h_0 = 1/\sqrt{2}$$

$$h_1 = 1/\sqrt{2}$$

- Daubechies D2 (two vanishing moments):

$$h_0 = 0.482\,962\,913\,145$$

$$h_1 = 0.836\,516\,303\,736$$

$$h_2 = 0.224\,143\,868\,042$$

$$h_3 = -0.129\,409\,522\,551$$

- Daubechies D3 (three vanishing moments):

$$h_0 = 0.332\,670\,552\,950$$

$$h_1 = 0.806\,891\,509\,311$$

$$h_2 = 0.459\,877\,502\,118$$

$$h_3 = -0.135\,011\,020\,010$$

$$h_4 = -0.085\,441\,273\,882$$

$$h_5 = 0.035\,226\,291\,882$$

- Coiflets C12 (four vanishing moments): the wavelets and the corresponding scaling function are shown in Figure 3.

Remarks The construction of orthogonal wavelets in $L^2(\mathbb{R})$ can be modified to obtain wavelets on the interval, that is, in $L^2([0, 1])$. Therewith, boundary wavelets are introduced, while in the interior of the interval the wavelets are not modified.

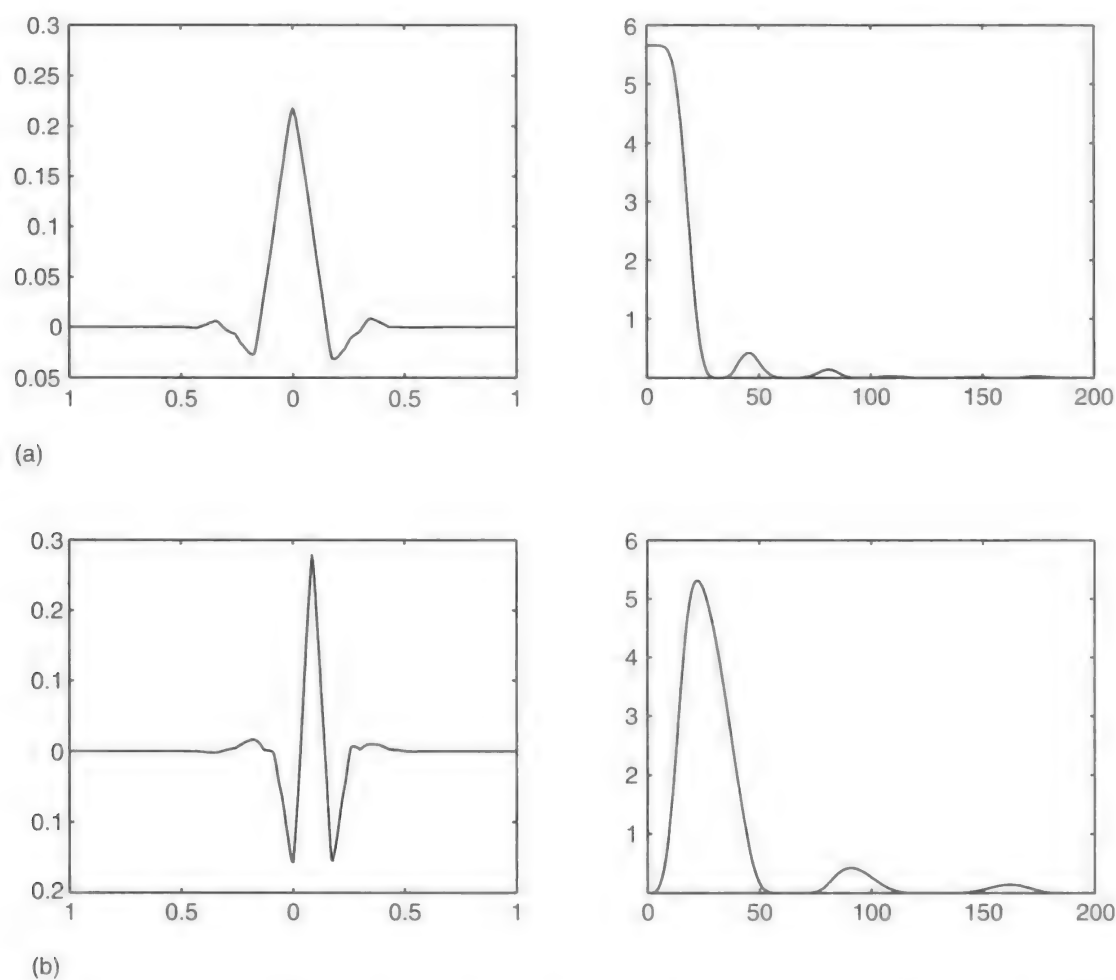


Figure 3 Orthogonal wavelets Coiflet C12. (a) Scaling function $\phi(x)$ (left) and $|\hat{\phi}(\omega)|$. (b) Wavelet $\psi(x)$ (left) and $|\hat{\psi}(\omega)|$.

A periodic MRA of $L^2(\mathbb{T})$, where $\mathbb{T}=\mathbb{R}/\mathbb{Z}$ denotes the torus, can also be constructed by periodizing the wavelets in $L^2(\mathbb{R})$, using

$$\psi^{\text{per}}(x)=\sum_{k\in\mathbb{Z}}\psi(x+k)$$

Relaxing the condition of orthogonality allows greater flexibility in the choice of the basis functions. For example, biorthogonal wavelets can be designed using different basis functions for analysis (^a) and synthesis (^s) which are related but no longer orthogonal. A couple of refinable scaling functions (ϕ^a,ϕ^s) with related wavelets (ψ^a,ψ^s) which are by construction biorthogonal generate a biorthogonal MRA V_j^a,V_j^s . From an algorithmic point of view, only two different filter couples (g^a,h^a) for the forward and (g^s,h^s) for the backward FWT are used, without changing the algorithm.

The multiresolution approach can be further generalized, for samplings on nonequidistant grids leading to the so-called second-generation wavelets.

Higher Dimensions

The previously presented one-dimensional construction can be extended to higher dimensions. For simplicity, we will consider only the two-dimensional case, since higher dimensions can be treated analogously.

Tensor product construction Having developed a one-dimensional orthonormal basis ψ_{ji} of $L^2(\mathbb{R})$, one could use these functions as building blocks in higher dimensions. One way of doing so is to take the tensor product of two one-dimensional bases and to define

$$\psi_{j_x,j_y,i_x,i_y}(x,y)=\psi_{j_x,i_x}(x)\psi_{j_y,i_y}(y) \tag{52}$$

The resulting functions constitute an orthonormal wavelet basis for $L^2(\mathbb{R}^2)$. Each function $f\in L^2(\mathbb{R}^2)$ can then be developed into

$$f(x,y)=\sum_{j_x,i_x}\sum_{j_y,i_y}\tilde{f}_{j_x,j_y,i_x,i_y}\psi_{j_x,j_y,i_x,i_y}(x,y) \tag{53}$$

with $\tilde{f}_{j_x,j_y,i_x,i_y}=\langle f,\psi_{j_x,j_y,i_x,i_y}\rangle$. However, in this basis the two variables x and y are dilatated separately

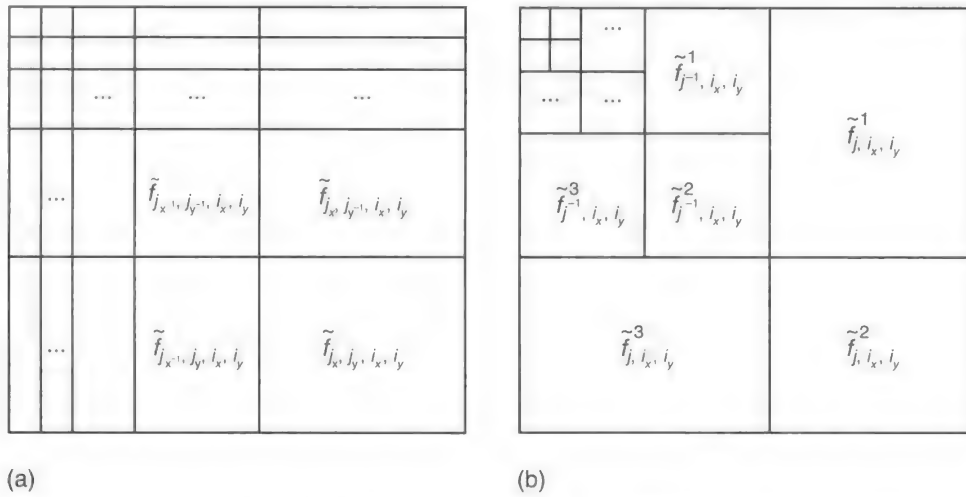


Figure 4a Schematic representation of the 2D (b) wavelet transforms: (a) Tensor product construction and (b) 2D MRA.

and therefore no longer form an MRA. This means that the functions ψ_{j_x, j_y} involve two scales, 2^{j_x} and 2^{j_y} , and each of the functions is essentially supported on a rectangle with these side-lengths. Hence, the decomposition is often called rectangular wavelet decomposition (cf. **Figure 4a**). From the algorithmic viewpoint, this is equivalent to applying the one-dimensional wavelet transform to the rows and the columns of a matrix or a function. For some applications, such a basis is advantageous, for others not. Often the notion of a scale has a certain meaning. For an application, one would like to have a unique scale assigned to each basis function.

Multiresolution construction Another much more interesting construction is the construction of a truly two-dimensional MRA of $L^2(\mathbb{R}^2)$. It can be obtained through the tensor product of two one-dimensional MRAs of $L^2(\mathbb{R})$. More precisely, one defines the spaces $V_j, j \in \mathbb{Z}$ by

$$V_j = V_j \otimes V_j \quad [54]$$

and $V_j = \overline{\text{span}}\{\phi_{j, i_x, i_y}(x, y) = \phi_{j, i_x}(x)\phi_{j, i_y}(y), i_x, i_y \in \mathbb{Z}\}$ fulfilling analogous properties as in the one-dimensional case.

Likewise, we define the complement space W_j to be the orthogonal complement of V_j in V_{j+1} , that is,

$$\begin{aligned} V_{j+1} &= V_{j+1} \otimes V_{j+1} \\ &= (V_j \oplus W_j) \otimes (V_j \oplus W_j) \end{aligned} \quad [55]$$

$$\begin{aligned} &= V_j \otimes V_j \oplus ((W_j \otimes V_j) \\ &\quad \oplus (V_j \otimes W_j) \oplus (W_j \otimes W_j)) \end{aligned} \quad [56]$$

$$= V_j \oplus W_j \quad [57]$$

It follows that the orthogonal complement $W_j = V_{j+1} \ominus V_j$ consists of three different types of functions and is generated by three different wavelets

$$\psi_{j, i_x, i_y}^\varepsilon(x, y) = \begin{cases} \psi_{j, i_x}(x)\phi_{j, i_y}(y), & \varepsilon = 1 \\ \phi_{j, i_x}(x)\psi_{j, i_y}(y), & \varepsilon = 2 \\ \psi_{j, i_x}(x)\psi_{j, i_y}(y), & \varepsilon = 3 \end{cases} \quad [58]$$

Observe that here the scale parameter j simultaneously controls the dilatation in x and y . We recall that in d dimensions this construction yields $2^d - 1$ types of wavelets spanning W_j .

Using [58], each function $f \in L^2(\mathbb{R}^2)$ can be developed into a multiresolution basis as

$$f(x, y) = \sum_j \sum_{i_x, i_y} \sum_{\varepsilon=1,2,3} \tilde{f}_{j, i_x, i_y}^\varepsilon \psi_{j, i_x, i_y}^\varepsilon(x, y) \quad [59]$$

with $\tilde{f}_{j, i_x, i_y}^\varepsilon = \langle f, \psi_{j, i_x, i_y}^\varepsilon \rangle$. A schematic representation of the wavelet coefficients is shown in **Figure 4b**. The algorithmic structure of the one-dimensional transforms carries over to the two-dimensional case by simple tensorization, that is, applying the filters at each decomposition step to rows and columns.

Remark The described two-dimensional wavelets and scaling functions are separable. This advantage is the ease of generation starting from one-dimensional MRAs. However, the main drawback of this construction is that three wavelets are needed to span the orthogonal complement space W_j . Another property should be mentioned. By construction, the wavelets are anisotropic, that is, horizontal, diagonal, and vertical directions are preferred.

Approximation Properties

Reproduction of Polynomials

A fundamental property of the MRA is the exact reproduction of polynomials. The vanishing moments of the wavelet ψ , that is, $\int_{\mathbb{R}} x^m \psi(x) dx = 0$

for $m=0,M-1$, is equivalent to the fact that polynomials up to degree $M-1$, can be expressed exactly as a linear combination of scaling functions, $p_m(x)=\sum_{n\in\mathbb{Z}}n^m\phi(x-n)$ for $m=0,M-1$. This so-called Strang–Fix condition proves that ψ has M vanishing moments if and only if any polynomial of degree $M-1$ can be written as a linear combination of scaling functions ϕ . Note that, as $p_m\notin L^2(\mathbb{R})$, the coefficients n^m are not in $\ell^2(\mathbb{Z})$.

Regularity and Local Decay of Wavelet Coefficients

The local or global regularity of a function is closely related to the decay of its wavelet coefficients. If a function is locally in $C^s(\mathbb{R})$ (the space of s -times continuously differentiable functions), it can be well approximated locally by a Taylor series of degree s . Consequently, its wavelet coefficients are small at fine scales, as long as the wavelet ψ has enough vanishing moments. The decay of the coefficients hence determines directly the error being made when truncating a wavelet sum at some scale.

Depending on the type of norm used and whether global or local characterization is concerned, various relations of this kind have been developed. Let us take as example the case of an α -Lipschitz function.

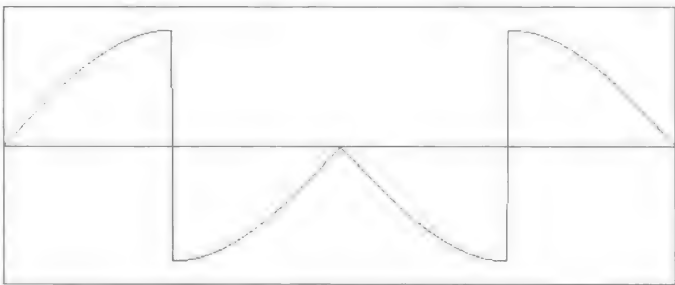
Suppose $f\in L^2(\mathbb{R})$, then for $[a,b]\subset\mathbb{R}$ the function f is α -Lipschitz with $0<\alpha<1$ for any $x_0\in[a,b]$, that is, $|f(x_0+h)-f(x_0)|\leq C|h|^\alpha$, if and only if there exists a constant A such that $|\tilde{f}_{ji}|\leq A2^{-j\alpha-1/2}$ for any (j,i) with $i/2^j\in[a,b]$.

This shows the relation between the local regularity of a function and the decay of its wavelet coefficients in scale.

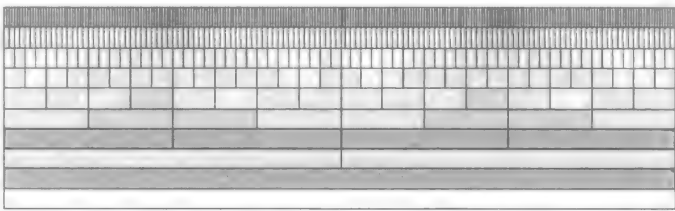
Example To illustrate the local decay of the wavelet coefficients, we consider in Figure 5 the function $f(x)=\sin(2\pi x)$ for $x\leq 1/4$ and $x\geq 3/4$ and $f(x)=-\sin(2\pi x)$ for $1/4<x<3/4$. The corresponding wavelet coefficients for quintic spline wavelets are plotted in logarithmic scale. The wavelet coefficients show that only in a local region around singularities the fine-scale coefficients are significant.

Linear Approximation

The exact reproduction of polynomials can be used to derive error estimates for the approximation of a function f at a given scale, which corresponds to linear approximation. We consider f belonging to the Sobolev space $W^{s,p}(\mathbb{R}^d)$, that is, the weak derivatives of f up to order s belong to $L^p(\mathbb{R}^d)$. The linear approximation of f at scale J , corresponding to the projection of f onto V_J , is then given by



(a)



(b)

Figure 5 Orthogonal wavelet decomposition using quintic spline wavelets: (a) function $f(x)=\sin(2\pi x)$ for $x\leq 1/4$ and $x\geq 3/4$ and $f(x)=-\sin(2\pi x)$ for $1/4<x<3/4$ sampled on a grid $x_i=i/2^J, i=0,\dots,2^J-1$ with $J=9$ and (b) corresponding wavelet coefficients $\log_{10}|\tilde{f}_{j,i}|$ for $i=0,\dots,2^j-1$ and $j=0,\dots,J-1$.

$$f_j(x)=\sum_{i=0}^{J-1}\sum_{i\in\mathbb{Z}}\tilde{f}_{j,i}\psi_{j,i}(x) \tag{60}$$

The approximation error can be estimated by

$$\|f-f_j\|_{L^p}<C2^{-J\min(s,m)/d} \tag{61}$$

where s denotes the smoothness of the function in L^p , d the space dimension, and m the number of vanishing moments of the wavelet ψ . In the case of poor global regularity of f , that is, for small s , a large number of scales J is needed to get a good approximation of f .

In Figure 6, we plot the linear approximation of the function f shown in Figure 5. The function f_6 is reconstructed using wavelet coefficients up to scale $J-1=5$, so that in total only 64 out of 512 coefficients are retained. We observe an oscillating behavior of f_j near the discontinuities of f which dominates the approximation error.

Nonlinear Approximation

Retaining the N largest wavelet coefficients in the wavelet expansion of f in [46], without imposing any *a priori* cutoff scale, yields the best N -term approximation f^N . In contrast to the linear approximation [60], it is called nonlinear approximation, since the choice of the retained coefficients depends

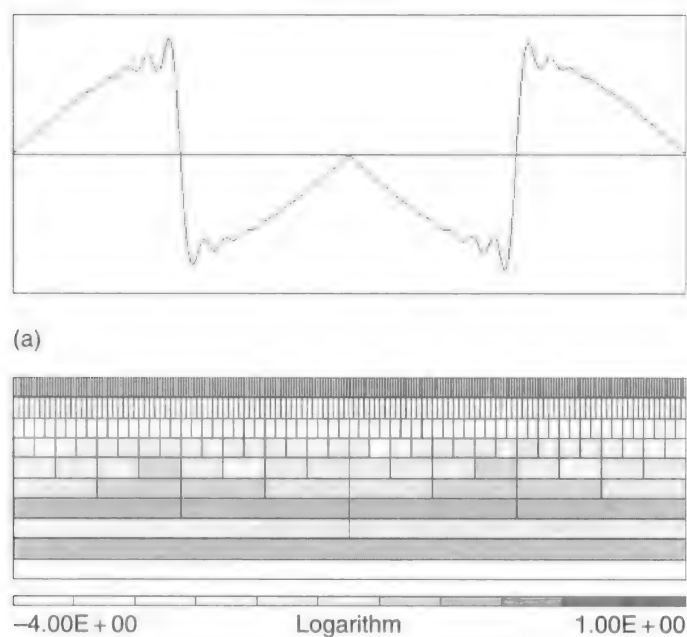


Figure 6 (a) Linear approximation f_J of the function f in **Figure 5** for $J=6$, reconstructed from 64 wavelet coefficients using quintic splines wavelets and (b) corresponding wavelet coefficients $\log_{10} |\tilde{f}_{j,i}|$ for $i=0, \dots, 2^j-1$ and $j=0, \dots, J-1$. Note that the coefficients for $J > 5$ have been set to zero.

on the function f . The mathematical theory has been formalized by Cohen, Dahmen, and De Vore.

The nonlinear approximation of the function f can then be written as

$$f^N(x) = \sum_{(j,i) \in \Lambda_N} \tilde{f}_{j,i} \psi_{j,i}(x) \tag{62}$$

where Λ_N denotes the ensemble of all multi-indices $\lambda=(j,i)$, indexing the N largest coefficients (measured in the l^p norm),

$$\Lambda_N = \{ \lambda_k, k=1, N \mid \| \tilde{f}_{\lambda_k} \|_{l^p} > \| \tilde{f}_{\mu} \|_{l^p} \quad \forall \mu \in \Lambda \} \tag{63}$$

with $\Lambda = \{ \mu=(j,i), j \geq 0, i \in \mathbb{Z} \}$. The nonlinear approximation leads to the following error estimate:

$$\| f - f^N \|_{L^p} < CN^{-s/d} \tag{64}$$

where s denotes the smoothness of f in the larger space $L^q(\mathbb{R}^d)$ with

$$\frac{1}{q} = \frac{1}{p} + \frac{s}{d}$$

which corresponds to the Sobolev embedding line (**Figure 7**). This estimate shows that the nonlinear approximation converges faster than the linear one, if f has a larger regularity in L^q , that is, $f \in W^{s,q}(\mathbb{R}^d)$, which is for example the case for functions with isolated singularities and for small q .

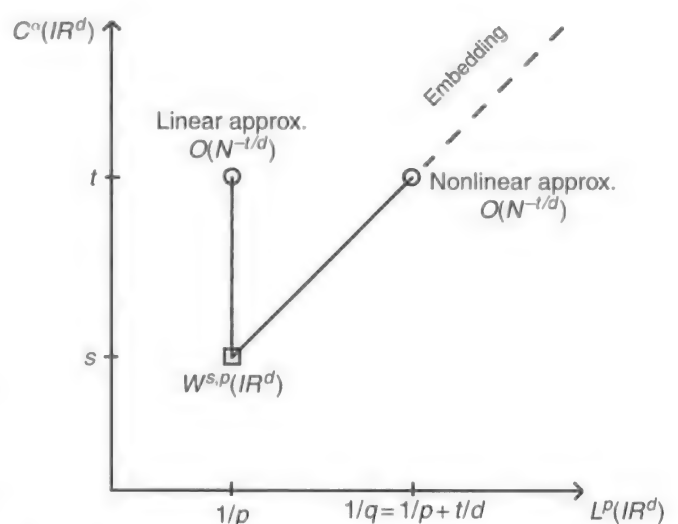


Figure 7 Schematic representation of linear and nonlinear approximation.

In **Figure 8**, we plot the nonlinear approximation of the function f shown in **Figure 5**. The function f^N is reconstructed using the strongest 64 wavelet coefficients out of 512 coefficients. Compared to the linear approximation (cf. **Figure 6**), the oscillations around the discontinuities disappear and the approximation error is reduced while using the same number of coefficients.

Compression and Preconditioning of Operators

The nonlinear approximation of functions can be extended to certain operators leading to an efficient

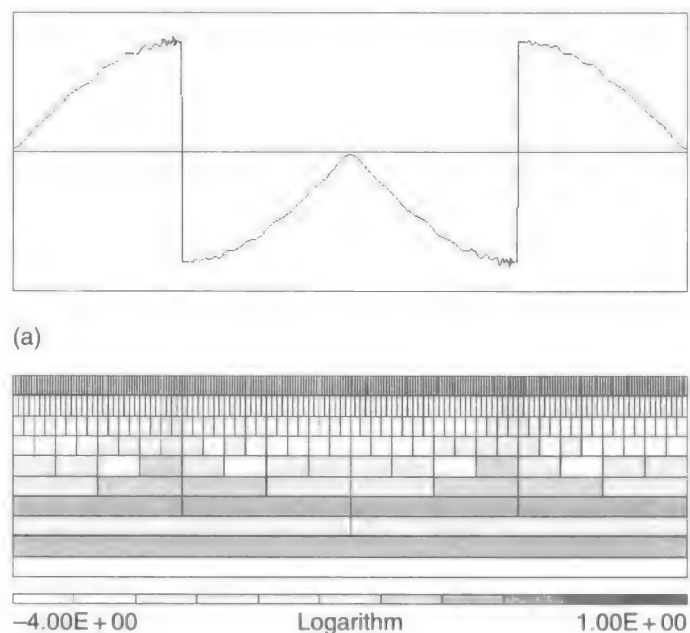


Figure 8 (a) Nonlinear approximation f^N of the function f in **Figure 5** reconstructed from the 64 largest wavelet coefficients using quintic splines wavelets, (b) retained wavelet coefficients $\log_{10} |\tilde{f}_{j,i}|$ for $i=0, \dots, 2^j-1$ and $j=0, \dots, J-1$.

representation in wavelet space, that is, to sparse matrices. For integral operators, for example, Calderon–Zygmund operators T on \mathbb{R} defined by

$$Tf(x) = \int_{\mathbb{R}} K(x, y) f(y) dy \quad [65]$$

where the kernel k satisfies

$$|k(x, y)| \leq \frac{C}{|x - y|}$$

and

$$\left| \frac{\partial}{\partial x} k(x, y) \right| + \left| \frac{\partial}{\partial y} k(x, y) \right| \leq \frac{C}{|x - y|^2}$$

their wavelet representation $\langle T\psi_{j,i}, \psi_{j',i'} \rangle$ is sparse and a large number of weak coefficients can be suppressed by simple thresholding of the matrix entries while controlling the precision. The resulting numerical scheme is called BCR algorithm and is due to Beylkin *et al.* (1991).

The characterization of function spaces by the decay of the wavelet coefficients and the corresponding norm equivalences can be used for diagonal preconditioning of integral or differential operators which leads to matrices with uniformly bounded condition numbers. For elliptic differential operators, for example, the Laplace operator ∇^2 the norm equivalence $\|\nabla^2 f\| \simeq \|2^{2j} \tilde{f}_{ji}\|$ can be used for preconditioning the matrix $\langle \nabla^2 \psi_{j,i}, \psi_{j',i'} \rangle$ by a simple diagonal scaling with 2^{-2j} to obtain a uniformly bounded condition number. For further details, we refer to the book of Cohen (2000).

Wavelet Denoising

We consider a function f which is corrupted by a Gaussian white noise $n \in \mathcal{N}(0, \sigma^2)$. The noise is spread over all wavelet coefficients \tilde{s}_λ , while, typically, the original function f is determined by only few significant wavelet coefficients. The aim is then to reconstruct the function f from the observed noisy signal $s = f + n$.

The principle of the wavelet denoising can be summarized in the following procedure:

- **Decomposition.** Compute the wavelet coefficients \tilde{s}_λ using the FWT.
- **Thresholding.** Apply the thresholding function ρ_ε to the wavelet coefficients \tilde{s}_λ , thus reducing the relative importance of the coefficients with small absolute value.
- **Reconstruction.** Reconstruct a denoised version s_C from the thresholded wavelet coefficients using the fast inverse wavelet transform.

The thresholding parameter ε depends on the variance of the noise and on the sample size N . The thresholding function ρ we consider corresponds to hard thresholding:

$$\rho_\varepsilon(a) = \begin{cases} a & \text{if } |a| > \varepsilon \\ 0 & \text{if } |a| \leq \varepsilon \end{cases} \quad [66]$$

Donoho and Johnstone (1994) have shown that there exists an optimal ε for which the relative quadratic error between the signal s and its estimator s_C is close to the minimax error for all signals $s \in \mathcal{H}$, where \mathcal{H} belongs to a wide class of function spaces, including Hölder and Besov spaces. They showed using the threshold

$$\varepsilon_D = \sigma_n \sqrt{2 \ln N} \quad [67]$$

yields an error which is close to the minimum error. The threshold ε_D depends only on the sampling N and on the variance of the noise σ_n ; hence, it is called universal threshold. However, in many applications, σ_n is unknown and has to be estimated from the available noisy data s . For this, the present authors have developed an iterative algorithm (see Azzolini *et al.* (2005)), which is sketched in the following:

1. **Initialization**
 - (a) given $s_k, k = 0, \dots, N - 1$. Set $i = 0$ and compute the FWT of s to obtain \tilde{s}_λ ;
 - (b) compute the variance σ_0^2 of s as a rough estimate of the variance of n and compute the corresponding threshold $\varepsilon_0 = (2 \ln N \sigma_0^2)^{1/2}$;
 - (c) set the number of coefficients considered as noise $N_{\text{noise}} = N$.
2. **Main loop repeat**
 - (a) set $N'_{\text{noise}} = N_{\text{noise}}$ and count the wavelet coefficients N_{noise} with modulus smaller than ε_i ;
 - (b) compute the new variance σ_{i+1}^2 from the wavelet coefficients whose modulus is smaller than ε_i and the new threshold $\varepsilon_{i+1} = (2(\ln N) \sigma_{i+1}^2)^{1/2}$;
 - (c) set $i = i + 1$ until $(N'_{\text{noise}} = N_{\text{noise}})$.
3. **Final step**
 - (a) compute s_C from the coefficients with modulus larger than ε_i using the inverse FWT.

Example To illustrate the properties of the denoising algorithm, we apply it to a one-dimensional test signal. We construct a noisy signal s by superposing a Gaussian white noise, with zero mean and variance $\sigma_W^2 = 1$, to a function f , normalized such that $((1/N) \sum_k |f_k|^2)^{1/2} = 10$. The number of samples is

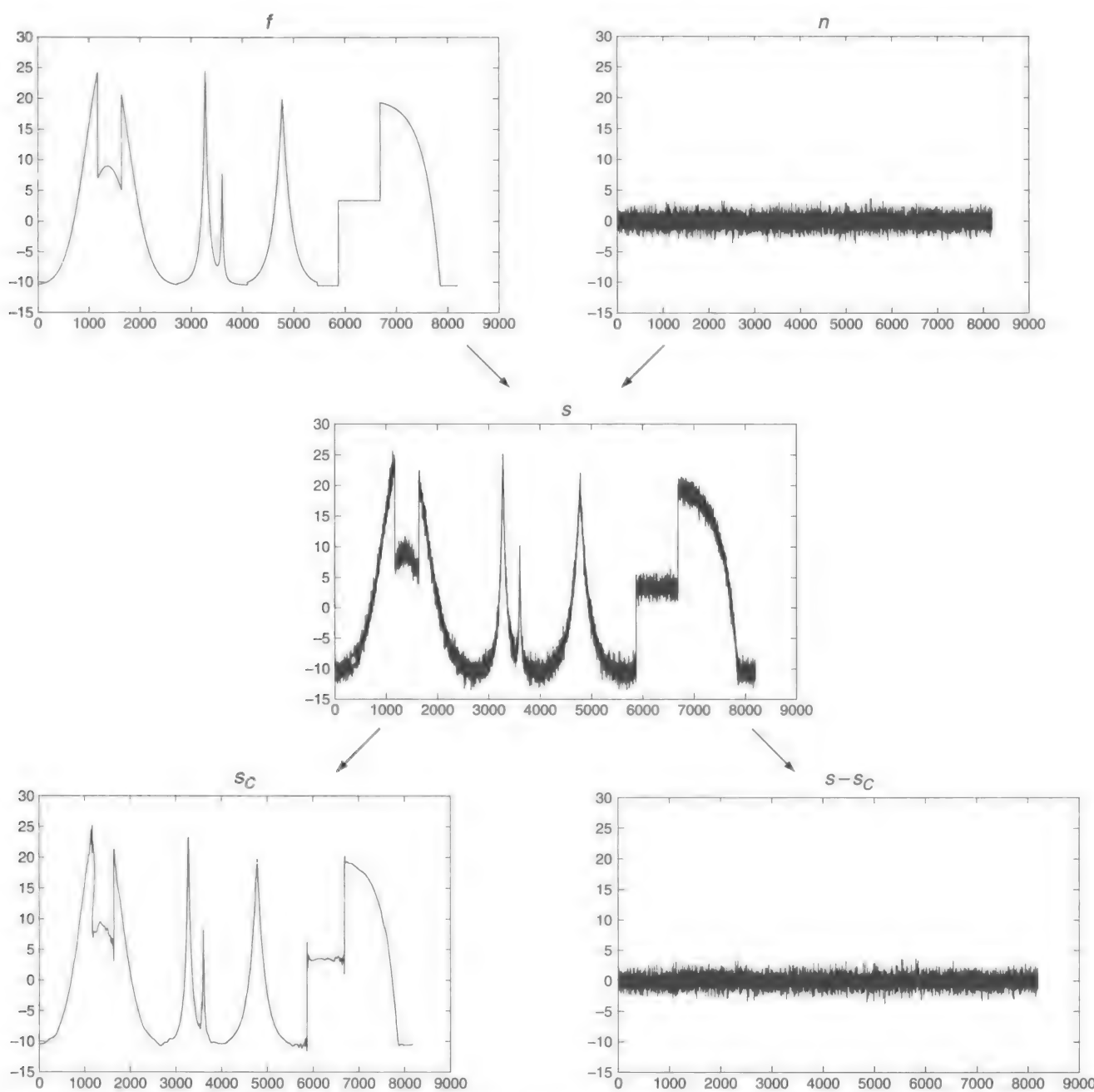


Figure 9 Construction (top) of a 1D noisy signal $s = f + n$ (middle), and results obtained by the recursive denoising algorithm (bottom).

$N = 8192$. Figure 9a shows the function f together with the noise n ; Figure 9b shows the constructed noisy signal s and Figure 9c shows the wavelet denoised signal s_C together with the extracted noise.

Acknowledgments

Marie Farge thankfully acknowledges Trinity College, Cambridge, UK, and CIRM, Marseille, France, for support while writing this paper. The authors also thank Barbara Burke for kindly revising their English.

See also: Coherent States; Fractal Dimensions in Dynamics; Homeomorphisms and Diffeomorphisms of

the Circle; Image Processing: Mathematics; Wavelets: Application to Turbulence; Wavelets: Applications.

Further Reading

Azzolini A, Farge M, and Schneider K (2005) Nonlinear wavelet thresholding: A recursive method to determine the optimal denoising threshold. *Applied and Computational Harmonic Analysis* 18(2): 177.

Beylkin, Coifman, and Rohklin (1991) Fast wavelet transforms and numerical algorithms. *Communications in Pure and Applied Mathematics* 44: 141.

Cohen A (2000) Wavelet methods in numerical analysis. In: Ciarlet PG and Lions JL (eds.) *Handbook of Numerical Analysis*, vol. 7. Amsterdam: Elsevier.

- Dahmen W (1997) Wavelets and multiscale methods for operator equations. *Acta Numerica* 6: 55–228.
- Daubechies I (1988) Orthonormal bases of compactly supported wavelets. *Communications in Pure and Applied Mathematics* 41: 909.
- Daubechies I (1992) *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM.
- Donoho and Johnstone (1994) Ideal spatial adaptation via wavelet shrinkage. *Biometrika* 81: 425.
- Grossmann A and Morlet J (1984) Decomposition of Hardy functions into square integrable wavelets of constant

- shape. *SIAM Journal of Mathematical Analysis*. 15(4): 723–736.
- Lemarié P-G and Meyer Yves (1986) Ondelettes et bases hilbertiennes. *Revista Matematica Iberoamericana* 2: 1.
- Mallat S (1989) Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Transactions of the American Mathematical Society* 315: 69.
- Mallat S (1998) *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic Press.

WDVV Equations and Frobenius Manifolds

B Dubrovin, SISSA-ISAS, Trieste, Italy

© 2006 Elsevier Ltd. All rights reserved.

Main Definition

WDVV equations of associativity (after E Witten, R Dijkgraaf, E Verlinde, and H Verlinde) is tantamount to the following problem: find a function $F(v)$ of n variables $v = (v^1, v^2, \dots, v^n)$ satisfying the conditions [1], [3], and [4] given below. First,

$$\frac{\partial^3 F(v)}{\partial v^1 \partial v^\alpha \partial v^\beta} \equiv \eta_{\alpha\beta} \quad [1]$$

must be a constant symmetric nondegenerate matrix. Denote $(\eta^{\alpha\beta}) = (\eta_{\alpha\beta})^{-1}$ the inverse matrix and introduce the functions

$$c_{\alpha\beta}^\gamma(v) = \eta^{\gamma\epsilon} \frac{\partial^3 F(v)}{\partial v^\epsilon \partial v^\alpha \partial v^\beta}, \quad \alpha, \beta, \gamma = 1, \dots, n \quad [2]$$

The main condition says that, for arbitrary v^1, \dots, v^n these functions must be structure constants of an associative algebra, that is, introducing a v -dependent multiplication law in the n -dimensional space by

$$a \cdot b := (c_{\alpha\beta}^1(v) a^\alpha b^\beta, \dots, c_{\alpha\beta}^n(v) a^\alpha b^\beta)$$

one obtains an n -parameter family of n -dimensional associative algebras (these algebras will automatically be also commutative). Spelling out this condition one obtains an overdetermined system of nonlinear PDEs for the function $F(v)$ often also called WDVV associativity equations

$$\begin{aligned} & \frac{\partial^3 F(v)}{\partial v^\alpha \partial v^\beta \partial v^\lambda} \eta^{\lambda\mu} \frac{\partial^3 F(v)}{\partial v^\mu \partial v^\gamma \partial v^\delta} \\ &= \frac{\partial^3 F(v)}{\partial v^\delta \partial v^\beta \partial v^\lambda} \eta^{\lambda\mu} \frac{\partial^3 F(v)}{\partial v^\mu \partial v^\gamma \partial v^\alpha} \end{aligned} \quad [3]$$

for arbitrary $1 \leq \alpha, \beta, \gamma, \delta \leq n$. (Summation over repeated indices will always be assumed.) The last one is the so-called quasihomogeneity condition

$$EF = (3-d)F + \frac{1}{2} A_{\alpha\beta} v^\alpha v^\beta + B_\alpha v^\alpha + C \quad [4]$$

where

$$E = \left(a_\beta^\alpha v^\beta + b^\alpha \right) \frac{\partial}{\partial v^\alpha}$$

for some constants a_β^α, b^α satisfying

$$a_1^\alpha = \delta_1^\alpha, \quad b^1 = 0$$

$A_{\alpha\beta}, B_\alpha, C, d$ are some constants. E is called Euler vector field and d is the charge of the Frobenius manifold.

For $n=1$ one has $F(v) = (1/6)v^3$. For $n=2$ one can choose

$$F(u, v) = \frac{1}{2} uv^2 + f(u)$$

only the quasihomogeneity [4] makes a constraint for $f(v)$. The first nontrivial case is for $n=3$. The solution to WDVV is expressed in terms of a function $f = f(x, y)$ in one of the two forms (in the examples all indices are written as lower):

$$\begin{aligned} d \neq 0: \quad & F = \frac{1}{2} v_1^2 v_3 + \frac{1}{2} v_1 v_2^2 + f(v_2, v_3) \\ & f_{xxy}^2 = f_{yyy} + f_{xxx} f_{xyy} \\ d = 0: \quad & F = \frac{1}{6} v_1^3 + v_1 v_2 v_3 + f(v_2, v_3) \\ & f_{xxx} f_{yyy} - f_{xxy} f_{xyy} = 1 \end{aligned} \quad [5]$$

The function $f(x, y)$ satisfies additional constraint imposed by [4]. Because of this the above PDEs [5] can be reduced (Dubrovin 1992, 1996) to a particular case of the Painlevé-VI equation (see Painlevé Equations).

The problem [1], [3], [4] is invariant with respect to linear changes of coordinates preserving the direction of the vector $\partial/\partial v^1$:

$$v^\alpha \mapsto \tilde{v}^\alpha = P_\beta^\alpha v^\beta + Q^\alpha, \quad \det(P_\beta^\alpha) \neq 0, \quad P_1^\alpha = \delta_1^\alpha$$

It is also allowed to add to $F(v)$ a polynomial of the degree at most 2. To consider more general non-linear changes of coordinates one has to give a coordinate-free form of the above equations [1], [3], [4]. This gives rise to the notion of Frobenius manifold introduced in Dubrovin (1992).

Recall that a Frobenius algebra is a pair (A, \langle, \rangle) , where A is a commutative associative algebra with a unity e over a field k (we will consider only the cases $k = \mathbb{R}, \mathbb{C}$) and \langle, \rangle is a k -bilinear symmetric non-degenerate invariant form on A , that is,

$$\langle x \cdot y, z \rangle = \langle x, y \cdot z \rangle$$

for arbitrary vectors x, y, z in A .

Definition Frobenius structure $(\cdot, e, \langle, \rangle, E, d)$ on the manifold M is a structure of a Frobenius algebra on the tangent spaces $T_v M = (A_v, \langle, \rangle_v)$ depending (smoothly, analytically, etc.) on the point $v \in M$. It must satisfy the following axioms.

FM1. The curvature of the metric \langle, \rangle_v on M (not necessarily positive definite) vanishes. Denote ∇ the Levi-Civita connection for the metric. The unity vector field e must be flat, $\nabla e = 0$.

FM2. Let c be the 3-tensor $c(x, y, z) := \langle x \cdot y, z \rangle$, $x, y, z \in T_v M$. The 4-tensor $(\nabla_w c)(x, y, z)$ must be symmetric in $x, y, z, w \in T_v M$.

FM3. A linear vector field $E \in \text{Vect}(M)$ (called Euler vector field) must be fixed on M , that is, $\nabla \nabla E = 0$, such that

$$\begin{aligned} \text{Lie}_E(x \cdot y) - \text{Lie}_E x \cdot y - x \cdot \text{Lie}_E y &= x \cdot y \\ \text{Lie}_E \langle, \rangle &= (2 - d) \langle, \rangle \end{aligned}$$

for some number $d \in k$ called “charge.”

The last condition (also called quasihomogeneity) means that the derivations $Q_{\text{Func}(M)} := E$, $Q_{\text{Vect}(M)} := \text{id} + \text{ad}_E$ define on the space $\text{Vect}(M)$ of vector fields on M a structure of graded Frobenius algebra over the graded ring of functions $\text{Func}(M)$.

Flatness of the metric \langle, \rangle implies local existence of a system of flat coordinates v^1, \dots, v^n on M . Usually, they are chosen in such a way that

$$e = \frac{\partial}{\partial v^1}$$

is the unity vector field. In such coordinates, the problem of local classification of Frobenius manifolds reduces to the WDVV associativity equations [1], [3], [4]. Namely, $\eta_{\alpha\beta}$ is the constant Gram matrix of the metric in these coordinates

$$\eta_{\alpha\beta} := \left\langle \frac{\partial}{\partial v^\alpha}, \frac{\partial}{\partial v^\beta} \right\rangle$$

The structure constants of the Frobenius algebra $A_v = T_v M$

$$\frac{\partial}{\partial v^\alpha} \cdot \frac{\partial}{\partial v^\beta} = c_{\alpha\beta}^\gamma(v) \frac{\partial}{\partial v^\gamma} \quad [6]$$

can be locally represented by third derivatives [2] of a function $F(v)$ satisfying [1], [3], [4]. The function $F(v)$ is called “potential” of the Frobenius manifold. It is defined up to adding of an at most quadratic polynomial in v^1, \dots, v^n .

A generalization of the above definition to the case of Frobenius supermanifolds can be found in Manin (1999). For the more general class of the so-called F -manifolds, the requirement of the existence of a flat invariant metric has been relaxed.

Deformed Flat Connection

One of the main geometrical structures of the theory of Frobenius manifolds is the deformed flat connection. This is a symmetric affine connection on $M \times \mathbb{C}^*$ defined by the following formulas:

$$\begin{aligned} \tilde{\nabla}_x y &= \nabla_x y + z x \cdot y, \quad x, y \in TM, z \in \mathbb{C}^* \\ \tilde{\nabla}_{d/dz} y &= \partial_z y + E \cdot y - \frac{1}{z} \mathcal{V} y \\ \tilde{\nabla}_x \frac{d}{dz} &= \tilde{\nabla}_{d/dz} \frac{d}{dz} = 0 \end{aligned} \quad [7]$$

where, as above, ∇ is the Levi-Civita connection for the metric \langle, \rangle and

$$\mathcal{V} := \frac{2-d}{2} E - \nabla E \quad [8]$$

is an operator on the tangent bundle TM antisymmetric with respect to \langle, \rangle ,

$$\langle \mathcal{V} x, y \rangle = - \langle x, \mathcal{V} y \rangle$$

Observe that the unity vector field e is an eigenvector of this operator with the eigenvalue

$$\mathcal{V} e = -\frac{d}{2} e$$

The connection $\tilde{\nabla} = \tilde{\nabla}(z)$ is not metric but it satisfies

$$\begin{aligned} \nabla \langle x, y \rangle &= \langle \tilde{\nabla}(-z)x, y \rangle + \langle x, \tilde{\nabla}(z)y \rangle \\ x, y &\in TM \end{aligned}$$

for any $z \in \mathbb{C}^*$. As it was discovered in Dubrovin (1992), vanishing of the curvature of the connection $\tilde{\nabla}$ is essentially equivalent to the axioms of Frobenius manifold.

Definition A “deformed flat function” $f(v;z)$ on a domain in $M \times \mathbb{C}^*$ is defined by the requirement of horizontality of the differential df

$$\tilde{\nabla} df = 0 \tag{9}$$

Due to vanishing of the curvature of $\tilde{\nabla}$ locally there exist n independent deformed flat functions $f_1(v;z), \dots, f_n(v;z)$ such that their differentials, together with the flat 1-form dz , span the cotangent plane $T^*_{(v;z)}(M \times \mathbb{C}^*)$. They will be called “deformed flat coordinates.” The global analytic properties of deformed flat coordinates can be derived, for the case of semisimple Frobenius manifolds, from the results of the section “Moduli of semisimple Frobenius manifolds” discussed later.

One can relax the definition of Frobenius manifold dropping the last axiom FM3. The potential $F(v)$ in this case satisfies [1] and [3] but not [4]. In this case, the deformed flat connection $\tilde{\nabla}$ is just a family of affine flat connections on M depending on the parameter $z \in \mathbb{C}$ given by the first line in [7]. The curvature and torsion of this family of connections vanishes identically in z . The deformed flat functions of $\tilde{\nabla}$ defined as in [9] can be chosen in the form of power series in z . The flatness equations written in the flat coordinates on M yield a recursion equation for the coefficients of these power series

$$\begin{aligned} \tilde{\nabla} df &= 0, \quad f = \sum_{p \geq 0} \theta_p(v) z^p \\ \partial_\lambda \partial_\mu f &= z c^\nu_{\lambda\mu}(v) \partial_\nu f \\ \partial_\lambda \partial_\mu \theta_0(v) &= 0 \\ \partial_\lambda \partial_\mu \theta_{p+1}(v) &= c^\nu_{\lambda\mu}(v) \partial_\nu \theta_p(v) \quad p \geq 0 \end{aligned} \tag{10}$$

Thus, $f(v;0)$ is just an affine linear function of the flat coordinates v^1, \dots, v^n ; the dependence on z can be considered as a deformation of the affine structure. This motivates the name “deformed flat coordinates.” The coefficients of the expansions of the deformed flat coordinates are the leading terms of the ε -expansion of the Hamiltonian densities of the integrable hierarchies associated with the Frobenius manifolds (see below).

Intersection Form of a Frobenius Manifold

Another important geometric structure on M is the intersection form of the Frobenius manifold. It is a symmetric bilinear form on the cotangent bundle T^*M defined by the formula

$$(\omega_1, \omega_2) = i_E \omega_1 \cdot \omega_2, \quad \omega_1, \omega_2 \in T^*M \tag{11}$$

Here the multiplication law on the cotangent planes is defined by means of the isomorphism

$$\langle, \rangle : TM \rightarrow T^*M$$

The discriminant $\Sigma \subset M$ is a proper analytic (for an analytic M) subset where the intersection form degenerates. One can introduce a new metric on the open subset $M \setminus \Sigma$ taking the inverse of the intersection form. A remarkable result of the theory of Frobenius manifolds is vanishing of the curvature of this new metric. Moreover, the new flat metric together with the following new multiplication:

$$x * y := x \cdot y \cdot E^{-1}$$

defines on $M \setminus \Sigma$ a structure of an almost-dual Frobenius manifold (Dubrovin 2004). In the original flat coordinates v^1, \dots, v^n the coordinate expressions for the new metric and for the associated Levi-Civita connection ∇^* , called the Gauss–Manin connection, read

$$\begin{aligned} g^{\alpha\beta}(v) &:= (dv^\alpha, dv^\beta) = E^\gamma(v) c^\alpha_\gamma{}^\beta(v) \\ \nabla^{*\alpha} dv^\beta &= \Gamma^{\alpha\beta}_\gamma(v) dv^\gamma \\ \Gamma^{\alpha\beta}_\gamma(v) &:= -g^{\alpha\nu}(v) \Gamma^\beta_{\nu\gamma}(v) = c^\alpha_\gamma{}^\epsilon(v) \left(\frac{1}{2} - \nu \right)^\beta_\epsilon \end{aligned} \tag{12}$$

The pair $(,)$ and \langle, \rangle of bilinear forms on T^*M possesses the following property crucial for understanding the relationships between Frobenius manifolds and integrable systems: they form a flat pencil. That means that on the complement to the subset

$$\Sigma_\lambda := \{v \in M \mid \det(g^{\alpha\beta}(v) - \lambda \eta^{\alpha\beta}) = 0\}$$

The inverse to the bilinear form

$$(,)_\lambda := (,) - \lambda \langle, \rangle \tag{13}$$

defines a metric with vanishing curvature. Flat functions $p = p(v; \lambda)$ for the flat metric are determined from the system

$$(\nabla^* - \lambda \nabla) dp = 0 \tag{14}$$

They are called “periods” of the Frobenius manifold. The periods $p(v; \lambda)$ are related to the deformed flat functions $f(v; z)$ by the suitably regularized Laplace-type integral transform

$$p(v; \lambda) = \int_0^\infty e^{-\lambda z} f(v; z) \frac{dz}{\sqrt{z}} \tag{15}$$

Choosing a system of n independent periods, one obtains a system of flat coordinates $p^1(v; \lambda), \dots, p^n(v; \lambda)$ for the metric $(,)_\lambda$ on $M \setminus \Sigma_\lambda$,

$$(dp^i(v; \lambda), dp^j(v; \lambda))_\lambda = G^{ij} \tag{16}$$

for some constant nondegenerate matrix G^{ij} .

The structure of a flat pencil on the Frobenius manifold M gives rise to a natural Poisson pencil (= bi-Hamiltonian structure) on the infinite-dimensional “manifold” $\mathcal{L}(M)$ consisting of smooth maps of a circle to M (the so-called loop space). In the flat coordinates v^1, \dots, v^n for the metric \langle, \rangle the Poisson pencil has the form

$$\begin{aligned} \{v^\alpha(x), v^\beta(y)\}_1 &= \eta^{\alpha\beta} \delta'(x-y) \\ \{v^\alpha(x), v^\beta(y)\}_2 &= g^{\alpha\beta}(v(x)) \delta'(x-y) \\ &\quad + \Gamma_\gamma^{\alpha\beta}(v(x)) v_x^\gamma \delta(x-y) \end{aligned} \quad [17]$$

By definition of the Poisson pencil, the linear combination $a_1\{, \}_1 + a_2\{, \}_2$ of the Poisson brackets is again a Poisson bracket for arbitrary constants a_1, a_2 . Choosing a system of n independent periods $p^i(v; \lambda), i=1, \dots, n$, as a new system of dependent variables, one obtains a reduction of the Poisson bracket $\{, \}_\lambda := \{, \}_2 - \lambda \{, \}_1$ for a given λ to the canonical form

$$\{p^i(v(x); \lambda), p^j(v(y); \lambda)\}_\lambda = G^{ij} \delta'(x-y) \quad [18]$$

Under an additional assumption of existence of tau function (Dubrovin 1996, Dubrovin and Zhang), one can prove that any Poisson pencil on $\mathcal{L}(M)$ of the form [17] with a nondegenerate matrix $(\eta^{\alpha\beta})$ comes from a Frobenius structure on M .

Canonical Coordinates on Semisimple Frobenius Manifolds

Definition The Frobenius manifold M is called semisimple if the algebras $T_\nu M$ are semisimple for ν belonging to an open dense subset in M .

Any n -dimensional semisimple Frobenius algebra over \mathbb{C} is isomorphic to the orthogonal direct sum of n copies of one-dimensional algebras. In this section, all the manifolds will be assumed to be complex analytic.

Near a semisimple point, the roots $u_i = u_i(\nu)$, $i=1, \dots, n$, of the characteristic equation

$$\det(g^{\alpha\beta}(\nu) - \lambda \eta^{\alpha\beta}) = 0 \quad [19]$$

can be used as local coordinates. The vectors $\partial/\partial u_i, i=1, \dots, n$, are basic idempotents of the algebras $T_\nu M$

$$\frac{\partial}{\partial u_i} \cdot \frac{\partial}{\partial u_j} = \delta_{ij} \frac{\partial}{\partial u_i}$$

We call u_1, \dots, u_n “canonical coordinates.” Observe that we violate the indices convention labeling the canonical coordinates by subscripts. We will never use summation over repeated indices when working

in the canonical coordinates. Actually, existence of canonical coordinates can be proved without using [4] (see details in Dubrovin (1992)).

Choosing locally branches of the square roots

$$\psi_{i1}(u) := \sqrt{\langle \partial/\partial u_i, \partial/\partial u_i \rangle}, \quad i=1, \dots, n \quad [20]$$

we obtain a transition matrix $\Psi = (\psi_{i\alpha}(u))$,

$$\frac{\partial}{\partial v^\alpha} = \sum_{i=1}^n \frac{\psi_{i\alpha}(u)}{\psi_{i1}(u)} \frac{\partial}{\partial u_i} \quad [21]$$

from the basis $\partial/\partial v^\alpha$ to the orthonormal basis

$$\begin{aligned} \langle f_i, f_j \rangle &= \delta_{ij} \\ f_1 &= \psi_{11}^{-1}(u) \frac{\partial}{\partial u_1} \\ f_2 &= \psi_{21}^{-1}(u) \frac{\partial}{\partial u_2}, \dots \\ f_n &= \psi_{n1}^{-1}(u) \frac{\partial}{\partial u_n} \end{aligned} \quad [22]$$

The matrix $\Psi(u)$ satisfies orthogonality condition

$$\Psi^*(u) \Psi(u) \equiv \eta, \quad \eta = (\eta_{\alpha\beta}), \quad \eta_{\alpha\beta} := \left\langle \frac{\partial}{\partial v^\alpha}, \frac{\partial}{\partial v^\beta} \right\rangle$$

In this formula Ψ^* stands for the transposed matrix. The lengths [20] coincide with the first column of this matrix.

Denote $V(u) = (V_{ij}(u))$ the matrix of the antisymmetric operator \mathcal{V} [8] with respect to the orthonormal frame

$$V(u) := \Psi(u) \mathcal{V} \Psi^{-1}(u) \quad [23]$$

The antisymmetric matrix $V(u) = (V_{ij}(u))$ satisfies the following system of commuting time-dependent Hamiltonian flows on the Lie algebra $\mathfrak{so}(n)$ equipped with the standard Lie–Poisson brackets $\{V_{ij}, V_{kl}\} = V_{il} \delta_{jk} - V_{jl} \delta_{ik} + V_{jk} \delta_{il} - V_{ik} \delta_{jl}$:

$$\frac{\partial V}{\partial u_i} = \{V, H_i(V; u)\}, \quad i=1, \dots, n \quad [24]$$

with quadratic Hamiltonians

$$H_i(V; u) = \frac{1}{2} \sum_{j \neq i} \frac{V_{ij}^2}{u_i - u_j} \quad [25]$$

The matrix $\Psi(u)$ satisfies

$$\begin{aligned} \frac{\partial \Psi}{\partial u_i} &= V_i(u) \Psi, \\ V_i(u) &:= \text{ad}_{E_i} \text{ad}_U^{-1}(V(u)), \quad i=1, \dots, n \end{aligned} \quad [26]$$

Here the matrix unity E_i has the entries $(E_i)_{ab} = \delta_{ai} \delta_{ib}$, $U = \text{diag}(u_1, \dots, u_n)$. Conversely, given a solution to [24] and [26], one can reconstruct the

Frobenius manifold structure by quadratures (Dubrovin 1998). The reconstruction depends on a choice of an eigenvector of the constant matrix $V = \Psi^{-1}(u) V(u) \Psi(u)$.

The system [24] coincides with the equations of isomonodromic deformations (see Isomonodromic Deformations) of the following linear differential operator with rational coefficients:

$$\frac{dY}{dz} = \left(U + \frac{V}{z} \right) Y \tag{27}$$

The latter is nothing but the last component of the deformed flat connection [7] written in the orthonormal frame [22]. Other components of the horizontality equations yield

$$\partial_i Y = (zE_i + V_i(u))Y, \quad i = 1, \dots, n \tag{28}$$

The compatibility conditions of the system [27] and [28] coincide with [24].

The integration of [24], [26] and, more generally, the reconstruction of the Frobenius structure can be reduced to a solution of a certain Riemann–Hilbert problem (see Riemann–Hilbert Problem).

The isomonodromic tau function of the semisimple Frobenius manifold is defined by

$$d \log \tau_I(u) = \sum_{i=1}^n H_i(V(u); u) du_i \tag{29}$$

It is an analytic function on a suitable unramified covering of the semisimple part of M .

Alternatively, eqns [24] can be represented as the isomonodromy deformations of the dual Fuchsian system

$$[U - \lambda] \frac{d\phi}{d\lambda} = \left(\frac{1}{2} + V \right) \tag{30}$$

The latter comes from the Gauss–Manin system for the periods $p = p(v; \lambda)$ of the Frobenius manifold written in the canonical coordinates [22].

Moduli of Semisimple Frobenius Manifolds

All n -dimensional semisimple Frobenius manifolds form a finite-dimensional space. They depend on $n(n - 1)/2$ essential parameters. To parametrize the Frobenius manifolds one can choose, for example, the initial data for the isomonodromy deformation equations [24]. Alternatively, they can be parametrized by monodromy data of the deformed flat connection according to the following construction.

The first part of the monodromy data is the spectrum $(V, <, >, \hat{\mu}, R)$ of the Frobenius manifold associated with the Poisson pencil. Here V is an

n -dimensional linear space equipped with a symmetric nondegenerate bilinear form $<, >$. Two linear operators on V , a semisimple operator $\hat{\mu}: V \rightarrow V$, and a nilpotent operator $R: V \rightarrow V$ must satisfy the following properties. First, the operator $\hat{\mu}$ is antisymmetric:

$$\hat{\mu}^* = -\hat{\mu} \tag{31}$$

and the operator R satisfies

$$R^* = -e^{-\pi i \hat{\mu}} R e^{\pi i \hat{\mu}} \tag{32}$$

Here the adjoint operators are defined with respect to the bilinear form $<, >$. The last condition to be imposed onto the operator R can be formulated in a simple way by choosing a basis e_1, \dots, e_n of eigenvectors of the semisimple operator $\hat{\mu}$,

$$\hat{\mu} e_\alpha = \mu_\alpha e_\alpha, \quad \alpha = 1, \dots, n$$

We require the existence of a decomposition

$$R = R_0 + R_1 + R_2 + \dots \tag{33}$$

where for any integer $k \geq 0$ the linear operator R_k satisfies

$$R_k e_\alpha \in \text{span}\{e_\beta \mid \mu_\beta = \mu_\alpha + k\} \quad \forall \alpha = 1, \dots, n \tag{34}$$

In the nonresonant case, such that none of the differences of the eigenvalues of $\hat{\mu}$ being equal to a positive integer, all the matrices R_1, R_2, \dots , are equal to zero. Observe a useful identity

$$z^\mu R z^{-\hat{\mu}} = R_0 + z R_1 + z^2 R_2 + \dots \tag{35}$$

More generally, for any operator $A: V \rightarrow V$ commuting with $e^{2\pi i \hat{\mu}}$ a decomposition is defined as

$$\begin{aligned} A &= \bigoplus_{k \in \mathbb{Z}} [A]_k \\ z^\mu A z^{-\hat{\mu}} &= \sum_{k \in \mathbb{Z}} z^k [A]_k \end{aligned} \tag{36}$$

In particular, $[R]_k = R_k, k \geq 0, [R]_k = 0, k < 0$.

One has to also choose an eigenvector e of the operator $\hat{\mu}$ such that $R_0 e = 0$; denote $-d/2$ the corresponding eigenvalue

$$e \in V, \quad \hat{\mu} e = -\frac{d}{2} e, \quad R_0 e = 0 \tag{37}$$

The second part of the monodromy data is a pair of linear operators

$$C: V \rightarrow \mathbb{C}^n, \quad S: \mathbb{C}^n \rightarrow \mathbb{C}^n$$

The space \mathbb{C}^n is assumed to be equipped with the standard complex Euclidean structure given by the sum of squares. The properties of the operators S, C depend on the choice of an unordered set

$u^0 = (u_1^0, \dots, u_n^0)$ of n pairwise distinct complex numbers and on a choice of a ray ℓ_+ on an auxiliary complex z -plane starting at the origin such that

$$\operatorname{Re} z(u_i^0 - u_j^0) \neq 0, \quad i \neq j, \quad z \in \ell_+ \quad [38]$$

Let us order the complex numbers in such a way that

$$e^{z(u_i^0 - u_j^0)} \rightarrow 0, \quad i < j, |z| \rightarrow \infty, \quad z \in \ell_+ \quad [39]$$

The operator S must be upper triangular

$$\begin{aligned} S &= (S_{ij}), \quad S_{ij} = 0, \quad i > j \\ S_{ii} &= 1, \quad i = 1, \dots, n \end{aligned} \quad [40]$$

The operator C must satisfy

$$C^* S C = e^{\pi i \hat{\mu}} e^{\pi i R} \quad [41]$$

Here the adjoint operator C^* is understood as follows:

$$C^*: \mathbb{C}^n \xrightarrow{\sim} \mathbb{C}^{n*} \rightarrow V^* \xrightarrow{<, >^{-1}} V$$

The group of diagonal $n \times n$ matrices

$$D = \operatorname{diag}(\pm 1, \dots, \pm 1)$$

acts on the pairs (S, C) by

$$S \mapsto D S D, \quad C \mapsto D C$$

One is to factor out the action of this diagonal group. Besides, the operator C is defined up to a left action of certain group of linear operators depending on the spectrum.

For the generic (i.e., nonresonant) case where $e^{2\pi i \hat{\mu}}$ has simple spectrum, the operator C is defined up to left multiplication by any matrix commuting with $e^{2\pi i \hat{\mu}}$. In this situation, the monodromy data $(\hat{\mu}, R, S, C)$ are locally uniquely determined by the $n(n-1)/2$ entries of the matrix S . Therefore, near a generic point, the variety of the monodromy data is a smooth manifold of the dimension $n(n-1)/2$. At nongeneric points, the variety can get additional strata.

The monodromy data S, C are determined at an arbitrary semisimple point of a Frobenius manifold in terms of the analytic properties of horizontal sections of the deformed flat connection $\tilde{\nabla}$ [7] in the complex z -plane (the so-called “Stokes matrix” and the “central connection matrix” of the operator [27]). Locally, they do not depend on the point of the semisimple Frobenius manifold (the isomonodromicity property).

We will now describe the reconstruction procedure giving a parametrization of semisimple Frobenius

manifolds in terms of the monodromy data $(\hat{\mu}, R, S, C)$.

Conversely, to reconstruct the Frobenius manifold near a semisimple point with the canonical coordinates u_1^0, \dots, u_n^0 , one is to solve the following boundary-value problem. Let

$$\ell = (-\ell_-) \cup \ell_+$$

be the oriented line on the complex z -plane chosen as in [38]. Here the ray ℓ_- is the opposite to ℓ_+ . Denote Π_R/Π_L the right/left half-planes with respect to ℓ . To reconstruct the Frobenius manifold, one is to find three matrix-valued functions $\Phi_0(z; u)$, $\Phi_R(z; u)$, and $\Phi_L(z; u)$:

$$\begin{aligned} \Phi_0(z; u) &: V \rightarrow \mathbb{C}^n \\ \Phi_{R/L}(z; u) &: \mathbb{C}^n \rightarrow \mathbb{C}^n \end{aligned}$$

for u close to u^0 such that $\Phi_0(z; u)$ is analytic and invertible for $z \in \mathbb{C}$, $\Phi_R(z; u)/\Phi_L(z; u)$ are analytic and invertible for $z \in \Pi_R/\Pi_L$ resp., and continuous up to the boundary $\ell \setminus 0$ and

$$\Phi_{R/L}(z; u) \sim 1 + O(1/z), \quad |z| \rightarrow \infty, \quad z \in \Pi_R/\Pi_L$$

The boundary values of the functions $\Phi_0(z; u)$, $\Phi_R(z; u)$, and $\Phi_L(z; u)$ must satisfy the following boundary-value problem (as above $U = \operatorname{diag}(u_1, \dots, u_n)$):

$$\Phi_R(z; u) = \Phi_L(z; u) e^{zU} S e^{-zU}, \quad z \in \ell_+ \quad [42]$$

$$\Phi_R(z; u) = \Phi_L(z; u) e^{zU} S^* e^{-zU}, \quad z \in \ell_- \quad [43]$$

$$\begin{aligned} \Phi_0(z; u) z^{\hat{\mu}} z^R &= \Phi_R(z; u) e^{zU} C, \quad z \in \Pi_R \\ \Phi_0(z; u) z^{\hat{\mu}} z^R &= \Phi_L(z; u) e^{zU} S C, \quad z \in \Pi_L \end{aligned} \quad [44]$$

Here $z^{\hat{\mu}} := e^{\hat{\mu} \log z}$, $z^R := e^{R \log z}$ are considered as $\operatorname{Aut}(V)$ -valued functions on the universal covering of $\mathbb{C} \setminus 0$; the branch cut in the definition of $\log z$ is chosen to be along ℓ_- .

The solution of the above boundary-value problem [42]–[44], if exists, is unique. It can be reduced to a certain Riemann–Hilbert problem, that is, to a problem of factorization of an analytic $n \times n$ nondegenerate matrix-valued function on the annulus

$$G(z; u), \quad r < |z| < R, \quad \det G(z; u) \neq 0$$

depending on the parameter $u = (u_1, \dots, u_n)$ in a product

$$G(z; u) = G_0(z; u)^{-1} G_\infty(z; u) \quad [45]$$

of two matrix-valued functions $G_0(z; u)$ and $G_\infty(z; u)$ analytic for $|z| < R$ and $r < |z| \leq \infty$ resp., with nowhere-vanishing determinant.

Existence of a solution to the Riemann–Hilbert problem for a given $u=(u_1,\ldots,u_n), u_i\neq u_j$ for $i\neq j$, means triviality of certain n -dimensional vector bundle over the Riemann sphere with the transition functions given by $G(z;u)$. Existence of the solution for $u=u^0$ implies solvability of the Riemann–Hilbert problem for u sufficiently close to u^0 . From these arguments, it can be deduced that the matrices $\Phi_0(z;u), \Phi_{R/L}(z;u)$ are analytic in $(z;u)$ for u sufficiently close to u^0 . Moreover, they can be analytically continued in u to the universal covering of the space of configurations of n distinct points on the complex plane:

$$(C^n\setminus\cup_{i\neq j}\{u_i=u_j\})/S_n\tag{46}$$

The resulting functions are meromorphic on the universal covering, according to the results of B Malgrange and T Miwa. The structure of the global analytic continuation is given (Dubrovin 1999) in terms of a certain action of the braid group

$$B_n=\pi_1((C^n\setminus\cup_{i\neq j}\{u_i=u_j\})/S_n)$$

on the monodromy data.

Examples of Frobenius Manifolds

Example 0 Trivial Frobenius manifold, $M=A_0$ a graded Frobenius algebra, $F(v)=(1/6)\langle e, v\cdot v\cdot v\rangle$ is a cubic polynomial.

First nontrivial examples appeared in the setting of 2D topological field theories (Dijkgraaf *et al.* 1991, Witten 1991) (see Topological Quantum Field Theory: Overview). Mathematical formalization of these ideas gives rise to the following two classes of examples.

Example 1 Frobenius structure on the base of an isolated hypersurface singularity. The construction (Hertling 2002, Sabbah 2002) uses the K Saito theory of periods of primitive forms. For the example of A_n singularity $f(x)=x^{n+1}$ the Frobenius structure on the base of universal unfolding

$$M_{A_n}=\{f_s(x)=x^{n+1}+s_1x^{n-1}+\cdots+s_n\mid s_1,\ldots,s_n\in\mathbb{C}\}$$

is constructed as follows (Dijkgraaf *et al.* 1991):

$$\begin{aligned} e &= \frac{\partial}{\partial s_n} \\ E &= \frac{1}{n+1}\sum (k+1)s_k\frac{\partial}{\partial s_k} \\ d &= \frac{n-1}{n+1} \end{aligned}$$

The multiplication is introduced by identifying the tangent space T_sM with the quotient algebra

$$T_sM_{A_n}=\mathbb{C}[x]/(f_s'(x))$$

The metric has the form

$$\langle\partial_{s_i},\partial_{s_j}\rangle=-(n+1)\operatorname{res}_{x=\infty}\frac{\partial f_s(x)/\partial s_i\partial f_s(x)/\partial s_j}{f_s'(x)}\,dx$$

The flat coordinates $v_\alpha=v_\alpha(s)$ can be found from the expansion of the solution to the equation $f_s(x)=k^{n+1}$,

$$x=k-\frac{1}{n+1}\left(\frac{v_n}{k}+\frac{v_{n-1}}{k^2}+\cdots+\frac{v_1}{k^n}\right)+O\left(\frac{1}{k^{n+2}}\right)$$

The potentials of the Frobenius manifolds M_{A_n} for $n=1,2,3$ read

$$\begin{aligned} F_{A_1} &= \tfrac{1}{6}v_1^3 \\ F_{A_2} &= \tfrac{1}{2}v_1^2v_2+\tfrac{1}{72}v_2^4 \\ F_{A_3} &= \tfrac{1}{2}v_1v_2^2+\tfrac{1}{2}v_1^2v_3+\tfrac{1}{16}v_2^2v_3^2+\tfrac{1}{960}v_3^5 \end{aligned}\tag{47}$$

The space of polynomials M_{A_n} can be identified with the orbit space of $\mathbb{C}/W(A_n)$ of the Weyl group of the type A_n . More generally (Dubrovin 1996), the orbit space $M_W:=\mathbb{C}^n/W$ of an arbitrary irreducible finite Coxeter group $W\subset O(n)$ carries a natural structure of a polynomial semisimple Frobenius manifold. Conversely, all irreducible polynomial semisimple Frobenius manifolds with positive degrees of the flat coordinates can be obtained by this construction (Hertling 2002). Generalizations for the orbit spaces of certain infinite groups were obtained in Dubrovin and Zhang (1998b) and Bertola (2000).

Example 2 Gromov–Witten (GW) invariants (see Topological Sigma Models). Let X be a smooth projective variety. We will assume for simplicity that $H^{\operatorname{odd}}(X)=0$. To every such variety, one can associate a bunch of rational numbers. They are expressed in terms of intersection theory of certain cycles on the moduli spaces $X_{g,m,\beta}$ of stable genus g and degree β curves on X with m marked points (see details in Kontsevich and Manin (1994)):

$$\begin{aligned} X_{g,m,\beta} &:= \{f:(C_g,x_1,\ldots,x_m)\rightarrow X, \\ &\quad f_*[C_g]=\beta\in H_2(X;\mathbb{Z})\} \end{aligned}\tag{48}$$

Denote $n:=\dim H^*(X;\mathbb{C})$. Choosing a basis $\phi_1=1, \phi_2,\ldots,\phi_n$ we define the numbers

$$\begin{aligned} &\langle\tau_{p_1}(\phi_{\alpha_1})\cdots\tau_{p_m}(\phi_{\alpha_m})\rangle_{g,\beta} \\ &:= \int_{[X_{g,m,\beta}]^{\operatorname{virt}}} ev_1^*(\phi_{\alpha_1})\wedge c_1^{p_1}(\mathcal{L}_1) \\ &\quad \wedge\cdots\wedge ev_m^*(\phi_{\alpha_m})\wedge c_1^{p_m}(\mathcal{L}_m) \end{aligned}\tag{49}$$

for arbitrary non-negative integers p_1, \dots, p_m . Here the evaluation maps $ev_i, i = 1, \dots, m$, are given by

$$ev_i: X_{g,m,\beta} \rightarrow X, \quad f \mapsto f(x_i)$$

The so-called tautological line bundles \mathcal{L}_i over $X_{g,m,\beta}$ by definition have the fiber $T_{x_i}^*C_g, i = 1, \dots, m$ (see the article Moduli Spaces: An Introduction regarding the construction of the so-called virtual fundamental class $[X_{g,m,\beta}]^{\text{virt}}$). The numbers [49] can be defined for an arbitrary compact symplectic manifold X where one is to deal with the intersection theory on the moduli spaces of pseudoholomorphic curves fixing a suitable almost-complex structure on X . They depend only on the symplectic structure on X . In particular, the numbers

$$<\tau_0(\phi_{\alpha_1}) \dots \tau_0(\phi_{\alpha_m})>_{g,\beta} \tag{50}$$

are called the genus g and degree β GW invariants of X . In certain cases, they admit an interpretation in terms of enumerative geometry of the variety X (Kontsevich and Manin 1994). The numbers [49] with some of $p_i > 0$ are called “gravitational descendents.”

One can form a generating functions of the numbers [49]

$$\mathcal{F}_g^X = \sum_m \sum_{\beta \in H_2(X;\mathbb{Z})} \frac{1}{m!} t^{\alpha_1,p_1} \dots t^{\alpha_m,p_m} <\tau_{p_1}(\phi_{\alpha_1}) \dots \tau_{p_m}(\phi_{\alpha_m})>_{g,\beta} \tag{51}$$

(summation over repeated indices $1 \leq \alpha_1, \dots, \alpha_m \leq n$ will always be assumed). Here $t^{\alpha,p}$ are indeterminates labeled by pairs (α,p) with $\alpha = 1, \dots, n, p = 0, 1, 2, \dots$. (Usually one is to insert in the definition of \mathcal{F}_g^X elements q^β of the Novikov ring $\mathbb{C}[H_2(X;\mathbb{Z})]$. However, due to the divisor axiom (Kontsevich and Manin 1994) and these insertions can be compensated by a suitable shift in the space of couplings $t = (t^{\alpha,p})$.) We finally introduce the full generating function called total GW potential (it is also called the free energy of the topological sigma model with the target space X)

$$\mathcal{F}^X(t;\epsilon) = \sum_{g \geq 0} \epsilon^{2g-2} \mathcal{F}_g^X \tag{52}$$

Restricting the genus-zero generating function onto the so-called small phase space

$$F^X(v) := \mathcal{F}_0^X(t^{\alpha,0} = v^\alpha, t^{\alpha,p>0} = 0) \tag{53}$$
$$v = (v^1, \dots, v^n)$$

one obtains a solution to the WDVV associativity equations. This solution defines a structure of

(formal) Frobenius manifold on $H^*(X)$ with the bilinear form η given by the Poincaré pairing

$$\eta_{\alpha\beta} = \int_X \phi_\alpha \wedge \phi_\beta$$

the unity

$$e = \frac{\partial}{\partial v^1}$$

and the Euler vector field

$$E = \sum_{\alpha=1}^n [(1 - q_\alpha)v^\alpha + r_\alpha] \frac{\partial}{\partial v^\alpha}$$

Here the numbers q_α, r_α are defined by the conditions

$$\phi_\alpha \in H^{2q_\alpha}(X), \quad c_1(X) = \sum_{\alpha} r_\alpha \phi_\alpha$$

The resulting Frobenius manifold will be denoted M_X . The corresponding n -parameter family of n -dimensional algebras on the tangent spaces $T_\nu M_X$ is also called “quantum cohomology” $QH^*(X)$. At the point $\nu_{cl} \in M_X$ of classical limit, the algebra $T_{\nu_{cl}} M_X$ coincides with the cohomology ring $H^*(X)$. In all known examples, the series [53] actually converges in a neighborhood of the point ν_{cl} . Therefore, one obtains a genuine Frobenius structure on a domain $M_X \subset H^*(X;\mathbb{C})/2\pi i H_2(X;\mathbb{Z})$. However, a general proof of convergence is still missing.

In particular, for $d = 1$, the quantum cohomology of complex projective line \mathbb{P}^1 is a two-dimensional Frobenius manifold with the potential, unity, and the Euler vector field

$$F(u,v) = \frac{1}{2}uv^2 + e^u, \\ e = \frac{\partial}{\partial v}, \\ E = v \frac{\partial}{\partial v} + 2 \frac{\partial}{\partial u}$$

For $d = 2$ one has a three-dimensional Frobenius manifold $QH^*(\mathbb{P}^2)$ with

$$F(v_1,v_2,v_3) = \frac{1}{2}v_1^2v_3 + \frac{1}{2}v_1v_2^2 + \sum_{k \geq 1} N_k \frac{v_3^{3k-1}}{(3k-1)!} e^{kv_2} \tag{54}$$
$$e = \frac{\partial}{\partial v_1} \\ E = v_1 \frac{\partial}{\partial v_1} + 3 \frac{\partial}{\partial v_2} - v_3 \frac{\partial}{\partial v_3}$$

where N_k = number of rational curves on \mathbb{P}^2 passing through $3k - 1$ generic points. WDVV [5] yields (Kontsevich and Manin 1994) recursion relations for

the numbers N_k starting from $N_1=1$. The closed analytic formula for the function [54] is still unknown.

Only for certain very exceptional X the Frobenius manifold M_X is semisimple (e.g., for $X=\mathbf{P}^d$). The general geometrical reasons of the semisimplicity of M_X are still to have been understood.

For the case $X=\text{Calabi-Yau manifold}$, the Frobenius manifold $QH^*(X)$ is never semisimple. This Frobenius structure can be computed in terms of the mirror symmetry construction (see Mirror Symmetry: A Geometric Survey).

Frobenius Manifold and Integrable Systems

The identities in the cohomology ring generated by the cocycles $ev_i^*(\phi_\alpha)$ and $\psi_i:=c_1(\mathcal{L}_i)$ can be recast into the form of differential equations for the generating function [52]. The variable $x:=t^{1,0}$ corresponding to $\phi_1=1$ plays a distinguished role in these differential equations. According to the idea of Witten (1991), the differential equations for the generating functions can be written as a hierarchy of systems of n evolutionary PDEs ($n=\dim H^*(X)$) for the unknown functions

w_alpha = <<tau_0(phi_alpha)tau_0(phi_1)>> = epsilon^2 (partial^2 F^X(t,epsilon) / (partial t^{1,0} partial t^{alpha,p})) [55]

The variable $x=t^{1,0}$ is the spatial variable of the equations of the hierarchy. The remaining parameters (coupling constants) $t^{\alpha,p}$ of the generating function play the role of the time variables. Witten suggested to use the two-point correlators

h_{alpha,p} = <<tau_{p+1}(phi_alpha)tau_0(phi_1)>> = epsilon^2 (partial^2 F^X(t,epsilon) / (partial t^{1,0} partial t^{alpha,p})) [56]

as the densities of the Hamiltonians of the flows of the hierarchy.

Existence of such a hierarchy can be proved for the case of GW invariants (and their descendents) of complex projective spaces \mathbf{P}^d (the results of Givental (2001) along with Dubrovin and Zhang (2005) can be used). For $d=0$ one obtains, according to the celebrated result by Kontsevich conjectured by Witten (see Topological Gravity, Two-Dimensional), the tau function of the solution to the KdV hierarchy (see Korteweg-de Vries Equation and Other Modulation Equations) specified by the initial condition,

u(x) |_{t=0} = x

For $d=1$ the hierarchy in question is the extended Toda lattice (see details in Dubrovin and Zhang (2004); see also Toda Lattices). For all other $d \geq 2$,

the needed integrable hierarchy is a new one. It can be associated (Dubrovin and Zhang) with an arbitrary n -dimensional semisimple Frobenius manifold M . The equations of the hierarchy have the form

w_i^j = A_j^i(w)w_x^j + epsilon^2 [B_j^i(w)w_{xxx}^j + C_{jk}^i(w)w_x^jw_{xx}^k + D_{jkl}^i(w)w_x^jw_x^kw_x^l] + O(epsilon^4), i=1,...,n [57]

The coefficients of ϵ^{2g} are graded homogeneous polynomials in u_x, u_{xx} , etc., of the degree $2g+1$,

deg d^m u / dx^m = m

The construction of the hierarchy is done in two steps. First, we construct the leading approximation (Dubrovin 1992). The equation of the hierarchy specifying the dependence on $t=t^{\alpha,p}$ at $\epsilon=0$ reads

(partial v / partial t^{alpha,p}) = partial_x (nabla theta_{alpha,p+1}(v)) alpha=1,...,n, p >= 0 [58]

The functions $\theta_{\alpha,p}(v), v \in M$, are the coefficients of expansion [10] of the deformed flat functions normalized by $\theta_{\alpha,0}=v_\alpha$. The solution $v=v(x,t)$ of interest is determined from the implicit function equations

v = xe + sum_{alpha,p} t^{alpha,p} nabla theta_{alpha,p}(v) [59]

Next, one has to find solution

Delta F = sum_{g >= 1} epsilon^{2g-2} F_g(v; v_x, ..., v^{(3g-2)}) [60]

of the following universal loop equation (closely related with the Virasoro conjecture of Eguchi and Xiong (1998)):

sum_{r >= 0} (partial Delta F / partial v^{gamma,r}) partial_x^r (1 / (E(v) - lambda))^gamma + sum_{r >= 1} (partial Delta F / partial v^{gamma,r}) sum_{k=1}^r (r choose k) partial_x^{k-1} partial_e p_alpha G^{alpha beta} partial_x^{r-k+1} partial^{gamma} p_beta = -1/16 tr (U - lambda)^{-2} + 1/4 tr [(U - lambda)^{-1} V]^2 + epsilon^2 / 2 sum ((partial^2 Delta F / (partial v^{gamma,k} partial v^{rho,l}) + (partial Delta F / partial v^{gamma,k}) (partial Delta F / partial v^{rho,l})) x partial_x^{k+1} partial^{gamma} p_alpha G^{alpha beta} partial_x^{l+1} partial^{rho} p_beta + epsilon^2 / 2 sum (partial Delta F / partial v^{gamma,k}) partial_x^{k+1} x [nabla (partial p_alpha(v; lambda) / partial lambda) . nabla (partial p_beta(v; lambda) / partial lambda) . v_x]^gamma G^{alpha beta} [61]

Here $\mathcal{U} = \mathcal{U}(v)$ is the operator of multiplication by $E(v)$, $p_\alpha = p_\alpha(v; \lambda)$, $\alpha = 1, \dots, n$, is a system of flat coordinates [16] of the bilinear form [13]. The substitution

$$v_\alpha \mapsto w_\alpha = v_\alpha + \epsilon^2 \partial_x \partial_{t^{\alpha,0}} \Delta \mathcal{F}(v; v_x, v_{xx}, \dots; \epsilon^2) \quad [62]$$

$$\alpha = 1, \dots, n$$

transforms [58] to [57]. The terms of the expansion [60] are not polynomial in the derivatives. For example (Dubrovin and Zhang 1998a),

$$\mathcal{F}_1 = \frac{1}{24} \sum_{i=1}^n \log u'_i + \log \frac{\tau_1(u)}{J^{1/24}(u)} \quad [63]$$

$$J(u) = \det \left(\frac{\partial v^\alpha}{\partial u_i} \right) = \pm \prod_{i=1}^n \psi_{i1}(u)$$

(the canonical coordinates have been used) where $\tau_1(u)$ is the isomonodromic tau function [29]. The transformation [62] applied to the solution [59] expresses higher-genus GW invariants of a variety X with semisimple quantum cohomology $QH^*(X)$ via the genus-zero invariants. For the particular case of $X = \mathbb{P}^2$, the formula [63] yields (Dubrovin and Zhang 1998a)

$$\frac{\phi''' - 27}{8(27 + 2\phi' - 3\phi'')} = -\frac{1}{8} + \sum_{k \geq 1} k N_k^{(1)} \frac{e^{kz}}{(3k)!}$$

Here

$$\phi(z) = \sum_{k \geq 0} N_k \frac{e^{kz}}{(3k-1)!}$$

is the generating function of the genus-zero GW invariants of \mathbb{P}^2 (see [54]) and $N_k^{(1)}$ = the number of elliptic plane curves of the degree k passing through $3k$ generic points.

See also: Bi-Hamiltonian Methods in Soliton Theory; Functional Equations and Integrable Systems; Integrable Systems: Overview; Isomonodromic Deformations; Korteweg–de Vries Equation and Other Modulation Equations; Mirror Symmetry: A Geometric Survey; Moduli Spaces: An Introduction; Painlevé Equations; Riemann–Hilbert Problem; Toda Lattices; Topological Gravity, Two-Dimensional; Topological Quantum Field Theory: Overview; Topological Sigma Models.

Further Reading

Bertola M (2000) Frobenius manifold structure on orbit space of Jacobi groups. I, II. *Differential Geometry and Applications* 13: 19–41, 213–233.

Dijkgraaf R, Verlinde H, and Verlinde E (1991) Topological strings in $d < 1$. *Nuclear Physics B* 352: 59–86.

Dubrovin B (1992) Integrable systems in topological field theory. *Nuclear Physics B* 379: 627–689.

Dubrovin B (1996) Geometry of 2D topological field theories. In: Francaviglia M and Greco S (eds.) *Integrable Systems and Quantum Groups, Montecatini Terme, 1993*, Springer Lecture Notes in Math., vol. 1620, pp. 120–348.

Dubrovin B (1998) Geometry and analytic theory of Frobenius manifolds. *Proceedings of the International Congress of Mathematicians*, vol. II (Berlin, 1998). *Doc. Math.* 1998, Extra Vol. II, 315–326.

Dubrovin B (1999) Painlevé transcendents in two-dimensional topological field theory. In: Conte R (ed.) *The Painlevé Property: 100 Years Later*, CRM Ser. Math. Phys. pp. 287–412. New York: Springer.

Dubrovin B (2004) On almost duality for Frobenius manifolds. In: Buchstaber VM and Krichever IM (eds.) *Geometry, Topology, and Mathematical Physics*, Amer. Math. Soc. Transl. Ser. 2, vol. 212, pp. 75–132. Providence, RI: American Mathematical Society.

Dubrovin B and Zhang Y (1998a) Bi-Hamiltonian hierarchies in 2D topological field theory at one-loop approximation. *Communications in Mathematical Physics* 198: 311–361.

Dubrovin B and Zhang Y (1998b) Extended affine Weyl groups and Frobenius manifolds. *Compositio Math.* 111: 167–219.

Dubrovin B and Zhang Y (2004) Virasoro symmetries of the extended Toda hierarchy. *Communications in Mathematical Physics* 250: 161–193.

Dubrovin B and Zhang Y (2001) Normal forms of hierarchies of integrable PDEs, Frobenius manifolds and Gromov–Witten invariants, math/0108160.

Dubrovin B and Zhang Y (2005) Integrable hierarchies of the topological type (to appear).

Eguchi T and Xiong CS (1998) Quantum cohomology at higher genus: topological recursion relations and Virasoro conditions. *Advances in Theoretical and Mathematical Physics* 2: 219–229.

Givental A (2001) Gromov–Witten invariants and quantization of quadratic Hamiltonians. *Moscow Mathematical Journal* 1: 551–568, 645.

Hertling C (2002) Frobenius Manifolds and Moduli Spaces for Singularities, *Cambridge Tracts in Mathematics*, vol. 151. Cambridge: Cambridge University Press.

Hitchin N (1997) Frobenius manifolds. With Notes by David Calderbank. In: *Gauge Theory and Symplectic Geometry (Montreal, PQ, 1995)*, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. vol. 488, pp. 69–112. Dordrecht: Kluwer Academic Publishers.

Kontsevich M and Manin Yu (1994) Gromov–Witten classes, quantum cohomology and enumerative geometry. *Communications in Mathematical Physics* 164: 525–562.

Manin Yu (1999) *Frobenius Manifolds, Quantum Cohomology, and Moduli Spaces*, American Mathematical Society Colloquium Publications, vol. 47. Providence, RI: American Mathematical Society.

Sabbah C (2002) Déformations isomonodromiques et variétés de Frobenius. *Savoirs Actuels (Les Ulis)*. Paris: CNRS Éditions.

Witten E (1991) Two-dimensional gravity and intersection theory on moduli space. *Surv. of Differential Geometry* 1: 243–310.

Weakly Coupled Oscillators

E M Izhikevich, The Neurosciences Institute,
San Diego, CA, USA
Y Kuramoto, Hokkaido University, Sapporo, Japan
© 2006 Elsevier Ltd. All rights reserved.

Introduction

Practically any physical, chemical, or biological system can exhibit rhythmic oscillatory activity, at least when the conditions are right. Winfree (2001) reviews the ubiquity of oscillations in nature, ranging from autocatalytic chemical reactions to pacemaker cells in the heart, to animal gates, and to circadian rhythms. When coupled, even weakly, oscillators interact via adjustment of their phases, that is, their timing, often leading to synchronization. In this chapter, we review the most important concepts needed to study and understand the dynamics of coupled oscillators.

From a mathematical point of view, an oscillator is a dynamical system,

$$\dot{x} = f(x), \quad x \in \mathbb{R}^m \tag{1}$$

having a limit-cycle attractor – periodic orbit $\gamma \subset \mathbb{R}^m$. Its period is the minimal $T > 0$ such that

$$\gamma(t) = \gamma(t + T) \quad \text{for any } t$$

and its frequency is $\Omega = 2\pi/T$. Let $x(0) = x_0 \in \gamma$ be an arbitrary point on the attractor, then the state of the system, $x(t)$, is uniquely defined by its phase $\vartheta \in S^1$ relative to x_0 , where S^1 is the unit circle.

Throughout this article, we assume that the periodic orbit γ is exponentially stable, which implies normal hyperbolicity. In this case, there is a continuous transformation $\Theta: U \rightarrow S^1$ defined in a neighborhood $U \supset \gamma$ such that $\vartheta(t) = \Theta(x(t))$ for any trajectory in U , that is, Θ maps solutions of [1] to solutions of

$$\dot{\vartheta} = \Omega \tag{2}$$

Such a transformation removes the amplitude but saves the phase of oscillation.

Accordingly, there is a continuous transformation that maps solutions of the weakly coupled network of n oscillators,

$$\dot{x}_i = f_i(x_i) + \varepsilon g_i(x_1, \dots, x_n, \varepsilon), \quad \varepsilon \ll 1 \tag{3}$$

onto solutions of the phase system

$$\dot{\vartheta}_i = \Omega_i + \varepsilon h_i(\vartheta_1, \dots, \vartheta_n, \varepsilon), \quad \vartheta_i \in S^1 \tag{4}$$

which is easier for studying the collective properties of [3].

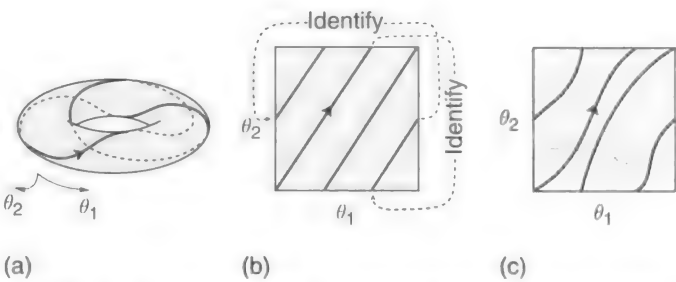


Figure 1 A 2-torus and its representation on the square. (Modified from Hoppensteadt and Izhikevich 1997.)

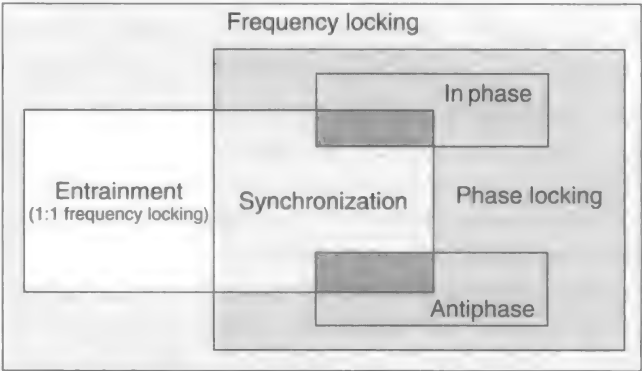


Figure 2 Various degrees of locking of oscillators. (Modified from Izhikevich 2006.)

The oscillators are said to be frequency locked when [4] has a stable periodic orbit $\vartheta(t) = (\vartheta_1(t), \dots, \vartheta_n(t))$ on the n -torus T^n , as in Figure 1a. The “rotation vector” or “winding ratio” of the orbit is the set of integers $q_1 : q_2 : \dots : q_n$ such that ϑ_1 makes q_1 rotations while ϑ_2 makes q_2 rotations, etc., as in the 2 : 3 frequency locking in Figure 1a. The oscillators are entrained when they are 1 : 1 : \dots : 1 frequency locked. The oscillators are phase locked when there is an $(n - 1) \times n$ integer matrix K having linearly independent rows such that $K\vartheta(t) = \text{const}$. For example, the two oscillators in Figure 1b are phase locked with $K = (2, 3)$, while those in Figure 1c are not. The oscillators are synchronized when they are entrained and phase locked. Synchronization is in-phase when $\vartheta_1(t) = \dots = \vartheta_n(t)$ and out-of-phase otherwise. Two oscillators are said to be synchronized antiphase when $\vartheta_1(t) - \vartheta_2(t) = \pi$. Frequency locking without phase locking, as in Figure 1c, is called phase trapping. The relationship between all these definitions is depicted in Figure 2.

Phase Resetting

An exponentially stable periodic orbit is a normally hyperbolic invariant manifold, hence its sufficiently small neighborhood, U , is invariantly foliated by

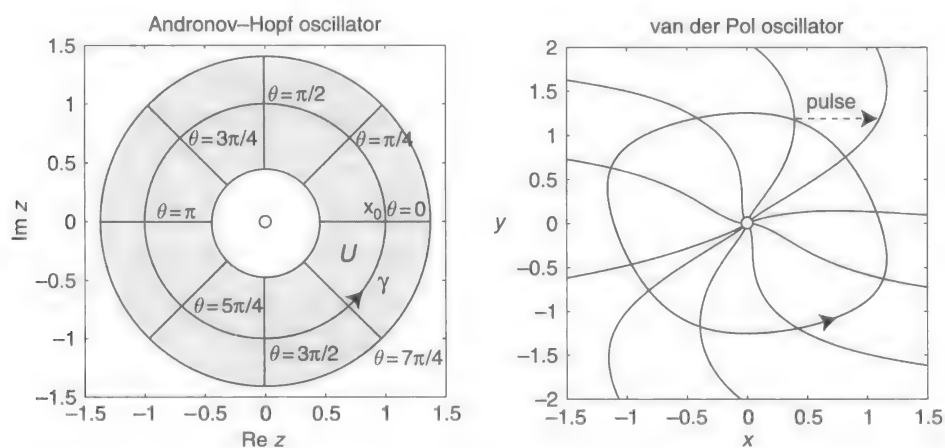


Figure 3 Isochrons of Andronov–Hopf oscillator ($\dot{z} = (1 + i)z - z|z|^2$, $z \in \mathbb{C}$) and van der Pol oscillator ($\dot{x} = x - x^3 - y$, $\dot{y} = x$).

stable submanifolds (Guckenheimer 1975) illustrated in Figure 3. The manifolds represent points having equal phases and, for this reason, they are called isochrons (from Greek “iso” meaning equal and “chronos” meaning time).

The geometry of isochrons determines how the oscillators react to perturbations. For example, the pulse in Figure 3, right, moves the trajectory from one isochron to another, thereby changing its phase. The magnitude of the phase shift depends on the amplitude and the exact timing of the stimulus relative to the phase of oscillation ϑ . Stimulating the oscillator at different phases, one can measure the phase transition curve (Winfree 2001)

$$\vartheta_{\text{new}} = \text{PTC}(\vartheta_{\text{old}})$$

and the phase resetting curve

$$\begin{aligned} \text{PRC}(\vartheta) &= \text{PTC}(\vartheta) - \vartheta \\ (\text{shift} &= \text{new phase} - \text{old phase}) \end{aligned}$$

Positive (negative) values of the PRC correspond to phase advances (delays). PRCs are convenient when the phase shifts are small, so that they can be magnified and clearly seen, as in Figure 4. PTCs are convenient when the phase shifts are large and comparable with the period of oscillation.

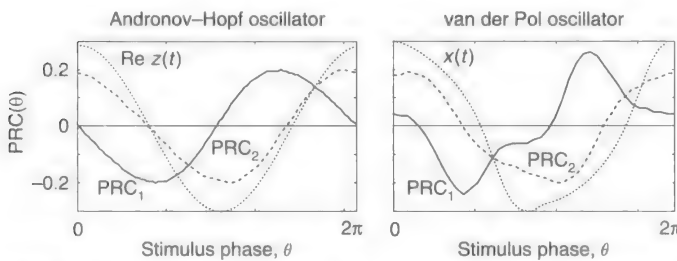


Figure 4 Examples of phase response curves (PRCs) of the oscillators in Figure 3. $\text{PRC}_1(\vartheta)$ and $\text{PRC}_2(\vartheta)$ correspond to horizontal (along the first variable) and vertical (along the second variable) pulses with amplitudes 0.2. An example of oscillation is plotted as a dotted curve in each subplot (not to scale).

In Figure 5 we depict phase portraits of the Andronov–Hopf oscillator receiving pulses of magnitude 0.5 (left) and 1.5 (right). Notice the drastic difference between the corresponding PRCs or PTCs. Winfree (2001) distinguishes two cases:

1. type 1 (weak) resetting results in continuous PRCs and PTCs with mean slope 1, and
2. type 0 (strong) resetting results in discontinuous PRCs and PTCs with mean slope 0.

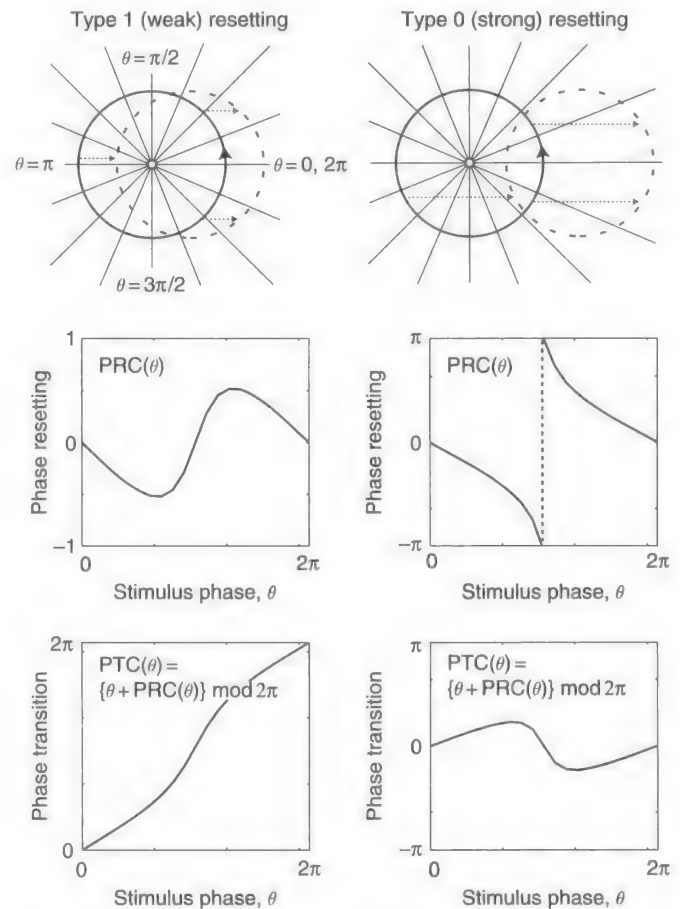


Figure 5 Types of phase resetting of the Andronov–Hopf oscillator in Figure 3.

The discontinuity of type 0 PRC in Figure 5 is a topological property that cannot be removed by reallocating the initial point x_0 that corresponds to zero phase. The discontinuity stems from the fact that the shifted image of the limit cycle (dashed circle) goes beyond the central equilibrium at which the phase is not defined.

The stroboscopic mapping of S^1 to itself, called Poincaré phase map,

$$\vartheta_{k+1} = \text{PTC}(\vartheta_k) \quad [5]$$

describes the response of an oscillator to a T -periodic pulse train. Here, ϑ_k denotes the phase of oscillation when the k th input pulse arrives. Its fixed points correspond to synchronized solutions, and its periodic orbits correspond to phase-locked states.

Weak Coupling

Now consider dynamical systems of the form

$$\dot{x} = f(x) + \varepsilon s(t) \quad [6]$$

describing periodic oscillators, $\dot{x} = f(x)$, forced by a weak time-dependent input $\varepsilon s(t)$, for example, from other oscillators in a network. Let $\Theta(x)$ denote the phase of oscillation at point $x \in U$, so that the map $\Theta: U \rightarrow S^1$ is constant along each isochron. This mapping transforms [6] into the phase model

$$\dot{\vartheta} = \Omega + \varepsilon Q(\vartheta) \cdot s(t)$$

with function $Q(\vartheta)$, illustrated in Figure 6, satisfying three equivalent conditions:

1. Winfree: $Q(\vartheta)$ is normalized PRC to infinitesimal pulsed perturbations;
2. Kuramoto: $Q(\vartheta) = \text{grad } \Theta(x)$; and
3. Malkin: Q is the solution to the adjoint problem

$$\dot{Q} = -\{Df(\gamma(t))\}^T Q \quad [7]$$

with the normalization $Q(t) \cdot f(\gamma(t)) = \Omega$ for any t .

The function $Q(\vartheta)$ can be found analytically in a few simple cases:

1. a nonlinear phase oscillator $\dot{x} = f(x)$ with $x \in S^1$ and $f > 0$ has $Q(\vartheta) = \Omega/f(\gamma(\vartheta))$;
2. a system near saddle-node on invariant circle bifurcation has $Q(\vartheta)$ proportional to $1 - \cos \vartheta$; and
3. a system near supercritical Andronov–Hopf bifurcation has $Q(\vartheta)$ proportional to $\sin(\vartheta - \psi)$, where $\psi \in S^1$ is a constant phase shift.

Other interesting cases, including homoclinic, relaxation, and bursting oscillators are considered by Izhikevich (2006).

Treating $s(t)$ in [6] as the input from the network, we can transform weakly coupled oscillators

$$\dot{x}_i = f_i(x_i) + \varepsilon \sum_{j=1}^n \overbrace{g_{ij}(x_i, x_j)}^{s_i(t)}, \quad x_i \in \mathbb{R}^m \quad [8]$$

to the phase model

$$\dot{\vartheta}_i = \Omega_i + \varepsilon Q_i(\vartheta_i) \cdot \sum_{j=1}^n \overbrace{g_{ij}(x_i(\vartheta_i), x_j(\vartheta_j))}^{s_i(t)} \quad [9]$$

having the form [4] with $h_i = Q_i \sum g_{ij}$, or the form

$$\dot{\vartheta}_i = \Omega_i + \varepsilon \sum_{j=1}^n h_{ij}(\vartheta_i, \vartheta_j)$$

where $h_{ij} = Q_i g_{ij}$. Introducing phase deviation variables $\vartheta_i = \Omega_i t + \varphi_i$, we transform this system into the form

$$\dot{\varphi}_i = \varepsilon \sum_{j=1}^n h_{ij}(\Omega_i t + \varphi_i, \Omega_j t + \varphi_j)$$

which can be averaged to

$$\dot{\varphi}_i = \varepsilon \sum_{j=1}^n H_{ij}(\varphi_i - \varphi_j) \quad [10]$$

with the functions

$$H_{ij}(\chi) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T h_{ij}(\Omega_i t, \Omega_j t - \chi) dt \quad [11]$$

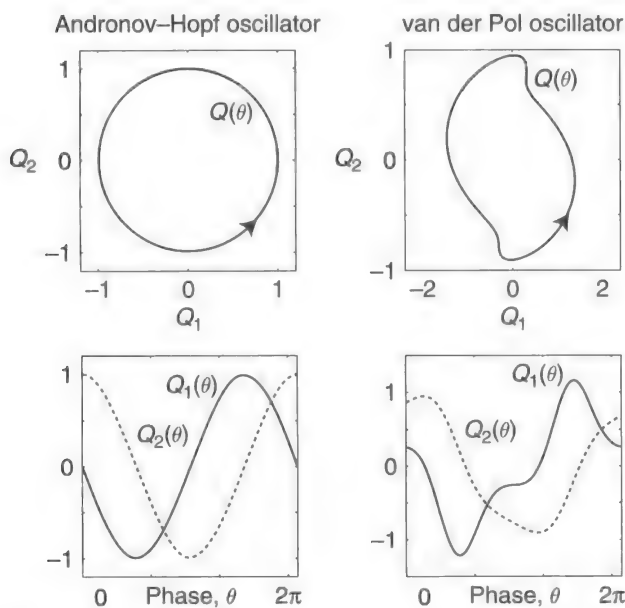


Figure 6 Solutions $Q = (Q_1, Q_2)$ to the adjoint problem [7] for oscillators in Figure 3.

describing the interaction between oscillators (Ermentrout and Kopell 1984). To summarize, we transformed weakly coupled system [8] into the phase model [10] with H given by [11] and each Q being the solution to the adjoint problem [7]. This constitutes the Malkin theorem for weakly coupled oscillators (Hoppensteadt and Izhikevich 1997, theorem 9.2).

Existence of one equilibrium of the phase model [10] implies the existence of the entire circular family of equilibria, since translation of all φ_i by a constant phase shift does not change the phase differences $\varphi_i - \varphi_j$ and hence the form of [10]. This family corresponds to a limit cycle of [8], on which all oscillators have equal frequencies and constant phase shifts, that is, they are synchronized, possibly out of phase.

We say that two oscillators, i and j , have resonant (or commensurable) frequencies when the ratio Ω_i/Ω_j is a rational number, for example, it is p/q for some integer p and q . They are nonresonant when the ratio is an irrational number. In this case, the function H_{ij} defined above is constant regardless of the details of the oscillatory dynamics or the details of the coupling, that is, dynamics of two coupled nonresonant oscillators is described by an uncoupled phase model. Apparently, such oscillators do not interact; that is, the phase of one of them cannot change the phase of the other one even on the long timescale of order $1/\varepsilon$.

Synchronization

Consider [8] with $n=2$, describing two mutually coupled oscillators. Let us introduce "slow" time $\tau = \varepsilon t$ and rewrite the corresponding phase model [10] in the form

$$\begin{aligned}\varphi_1' &= \omega_1 + H_{12}(\varphi_1 - \varphi_2) \\ \varphi_2' &= \omega_2 + H_{21}(\varphi_2 - \varphi_1)\end{aligned}$$

where $' = d/d\tau$ and $\omega_i = H_{ii}(0)$ is the frequency deviation from the natural oscillation, $i=1, 2$. Let $\chi = \varphi_2 - \varphi_1$ denote the phase difference between the oscillators; then

$$\chi' = \omega + H(\chi) \quad [12]$$

where

$$\omega = \omega_2 - \omega_1 \text{ and } H(\chi) = H_{21}(\chi) - H_{12}(-\chi)$$

is the frequency mismatch and the antisymmetric part of the coupling, respectively, illustrated in Figure 7, dashed curves. A stable equilibrium of [12] corresponds to a stable limit cycle of the phase model.

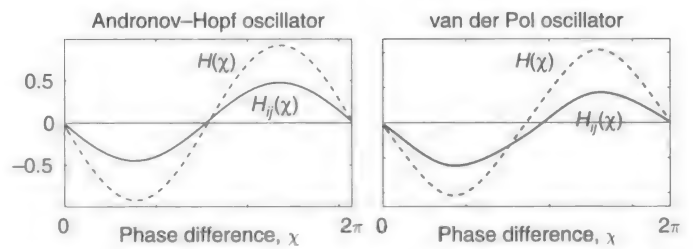


Figure 7 Solid curves: functions $H_{ij}(\chi)$ defined by [11] corresponding to the gap-junction input $g(x_i, x_j) = (x_{j1} - x_{i1}, 0)$. Dashed curves: functions $H(\chi) = H_{ij}(\chi) - H_{ij}(-\chi)$. Parameters are as in Figure 3.

All equilibria of [12] are solutions to $H(\chi) = -\omega$, and they are intersections of the horizontal line $-\omega$ with the graph of H . They are stable if the slope of the graph is negative at the intersection. If oscillators are identical, then $H(\chi)$ is an odd function (i.e., $H(-\chi) = -H(\chi)$), and $\chi=0$ and $\chi=\pi$ are always equilibria, possibly unstable, corresponding to the in-phase and antiphase synchronized solutions. The in-phase synchronization of gap-junction coupled oscillators in Figure 7 is stable because the slope of H (dashed curves) is negative at $\chi=0$. The max and min values of the function H determine the tolerance of the network to the frequency mismatch ω , since there are no equilibria outside this range.

Now consider a network of $n > 2$ weakly coupled oscillators [8]. To determine the existence and stability of synchronized states in the network, we need to study equilibria of the corresponding phase model [10]. The vector $\phi = (\phi_1, \dots, \phi_n)$ is an equilibrium of [10] when

$$0 = \omega_i + \sum_{j \neq i}^n H_{ij}(\phi_i - \phi_j) \quad (\text{for all } i) \quad [13]$$

It is stable when all eigenvalues of the linearization matrix (Jacobian) at ϕ have negative real parts, except one zero eigenvalue corresponding to the eigenvector along the circular family of equilibria (ϕ plus a phase shift is a solution of [13] too since the phase shifts $\phi_j - \phi_i$ are not affected).

In general, determining the stability of equilibria is a difficult problem. Ermentrout (1992) found a simple sufficient condition. If

1. $a_{ij} = H'_{ij}(\phi_i - \phi_j) \leq 0$, and
2. the directed graph defined by the matrix $a = (a_{ij})$ is connected, (i.e., each oscillator is influenced, possibly indirectly, by every other oscillator),

then the equilibrium ϕ is neutrally stable, and the corresponding limit cycle $x(t + \phi)$ of [8] is asymptotically stable.

Another sufficient condition was found by Hoppensteadt and Izhikevich (1997). If system [10] satisfies

1. $\omega_1 = \cdots = \omega_n = \omega$ (identical frequencies)
2. $H_{ij}(-\chi) = -H_{ji}(\chi)$ (pairwise odd coupling)

for all i and j , then the network dynamics converge to a limit cycle. On the cycle, all oscillators have equal frequencies $1 + \varepsilon\omega$ and constant phase deviations.

The proof follows from the observation that [10] is a gradient system in the rotating coordinates $\varphi = \omega\tau + \phi$ with the energy function

$$E(\phi) = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n R_{ij}(\phi_i - \phi_j)$$

where

$$R_{ij}(\chi) = \int_0^\chi H_{ij}(s) \, ds$$

One can check that $dE(\phi)/d\tau = -\sum (\phi_i')^2 \leq 0$ along the trajectories of [12] with equality only at equilibria.

Mean-Field Approximations

Let us represent the phase model [10] in the form

$$\varphi_i' = \omega_i + \sum_{j \neq i}^n H_{ij}(\varphi_i - \varphi_j)$$

where $' = d/d\tau$, $\tau = \varepsilon t$ is the slow time, and $\omega_i = H_{ii}(0)$ are random frequency deviations. Collective dynamics of this system can be analyzed in the limit $n \rightarrow \infty$. We illustrate the theory using the special case, $H(\chi) = -\sin \chi$, known as the Kuramoto (1984) model:

$$\varphi_i' = \omega_i + \frac{K}{n} \sum_{j=1}^n \sin(\varphi_j - \varphi_i), \quad \varphi_i \in [0, 2\pi] \quad [14]$$

where $K > 0$ is the coupling strength and the factor $1/n$ ensures that the model behaves well as $n \rightarrow \infty$. The complex-valued sum of all phases,

$$r e^{i\psi} = \frac{1}{n} \sum_{j=1}^n e^{i\varphi_j}$$

(Kuramoto synchronization index) [15]

describes the degree of synchronization in the network. Apparently, the in-phase synchronized state $\varphi_1 = \cdots = \varphi_n$ corresponds to $r=1$ with ψ being the population phase. In contrast, the incoherent state with all φ_i having different values

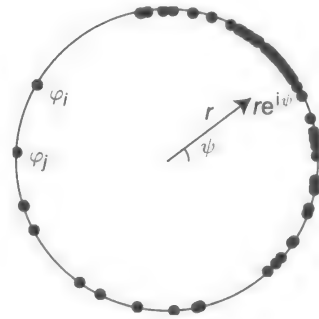


Figure 8 Kuramoto synchronization index [15] describes the degree of coherence in the network [14].

randomly distributed on the unit circle, corresponds to $r \approx 0$. Intermediate values of r correspond to a partially synchronized or coherent state, depicted in Figure 8. Some phases are synchronized forming a cluster, while others roam around the circle.

Multiplying both sides of [15] by $e^{-i\varphi_i}$ and considering only the imaginary parts, we can rewrite [14] in the equivalent form

$$\varphi_i' = \omega_i + Kr \sin(\psi - \varphi_i)$$

that emphasizes the mean-field character of interactions between the oscillators: they all are pulled into the synchronized cluster ($\varphi_i \rightarrow \psi$) with the effective strength proportional to the cluster size r . This pull is offset by the random frequency deviations ω_i that pull away from the cluster.

Let us assume that ω_i 's are distributed randomly around 0 with a symmetrical probability density function $g(\omega)$, for example, Gaussian. Kuramoto has shown that in the limit $n \rightarrow \infty$, the cluster size r obeys the self-consistency equation

$$r = rK \int_{-\pi/2}^{+\pi/2} g(Kr \sin \varphi) \cos^2 \varphi \, d\varphi \quad [16]$$

Notice that $r=0$, corresponding to the incoherent state, is always a solution of this equation. When the coupling strength K is greater than a certain critical value,

$$K_c = \frac{2}{\pi g(0)}$$

an additional, nontrivial solution $r > 0$ appears, which corresponds to a partially synchronized state. Expanding g in a Taylor series, one gets the scaling $r = \sqrt{16(K - K_c)/(-g''(0)\pi K_c^4)}$. Thus, the stronger the coupling K relative to the random distribution of frequencies, the more oscillators synchronize into a coherent cluster. The issue of stability of incoherent and partially synchronized states is discussed by Strogatz (2000). Other generalizations of the Kuramoto model are reviewed by Acebron *et al.* (2005). An extended version of this article with the

emphasis on computational neuroscience can be found in the recent book by Izhikevich (2006).

See also: Bifurcations of Periodic Orbits; Dynamical Systems and Thermodynamics; Hamiltonian Systems: Stability and Instability Theory; Singularity and Bifurcation Theory; Stability Theory and KAM; Synchronization of Chaos.

Further Reading

- Acebron JA, Bonilla LL, Vicente CJP, Ritort F, and Spigler R (2005) The Kuramoto model: a simple paradigm for synchronization phenomena. *Reviews of Modern Physics* 77: 137–186.
- Ermentrout GB (1992) Stable periodic solutions to discrete and continuum arrays of weakly coupled nonlinear oscillators. *SIAM Journal on Applied Mathematics* 52: 1665–1687.
- Ermentrout GB and Kopell N (1984) Frequency plateaus in a chain of weakly coupled oscillators, I. *SIAM Journal on Applied Mathematics* 15: 215–237.

- Glass L and MacKey MC (1988) *From Clocks to Chaos*. Princeton: Princeton University Press.
- Guckenheimer J (1975) Isochrons and phaseless sets. *Journal of Mathematical Biology* 1: 259–273.
- Hoppensteadt FC and Izhikevich EM (1997) *Weakly Connected Neural Networks*. New York: Springer.
- Izhikevich EM (1999) Weakly connected quasiperiodic oscillators, FM interactions, and multiplexing in the brain. *SIAM Journal on Applied Mathematics* 59: 2193–2223.
- Izhikevich EM (2006) *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. Cambridge, MA: The MIT Press.
- Kuramoto Y (1984) *Chemical Oscillations, Waves, and Turbulence*. New York: Springer.
- Pikovsky A, Rosenblum M, and Kurths J (2001) *Synchronization: A Universal Concept in Nonlinear Science*. Cambridge: Cambridge University Press.
- Strogatz SH (2000) From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D* 143: 1–20.
- Winfree A (2001) *The Geometry of Biological Time*, 2nd edn. New York: Springer.

Wheeler-De Witt Theory

J Maharana, Institute of Physics, Bhubaneswar, India

© 2006 Elsevier Ltd. All rights reserved.

Introduction

It is recognized that one of the outstanding problems in modern physics is to formulate the quantum theory of gravity, synthesizing the principles of quantum mechanics and general theory of relativity. The fundamental units for measuring time, length, and energy, known as Planck time, Planck length, and Planck energy, respectively, are defined to be $t_{\text{Pl}} = (\hbar G/c^5)^{1/2} = 5.39 \times 10^{-44}$ s, $l_{\text{Pl}} = (\hbar G/c^3)^{1/2} = 1.61 \times 10^{-33}$ cm, and $m_{\text{Pl}} = (\hbar c/G)^{1/2} = 2.17 \times 10^{-5}$ g, in terms of the Newton's constant, G , velocity of light, c , and $\hbar = h/2\pi$, h being the Planck's constant. We may conclude, on dimensional arguments, that quantum gravity effects will play an important role when we consider physical phenomena in the vicinity of these scales. Therefore, when we probe very short distances, consider collisions at Planckian energies, and envisage evolution of the universe in the Planck era, the quantum gravity will come into play in a predominant manner. The purpose of this article is to present an overview of an approach to quantize Einstein's theory of gravity, pioneered by Wheeler and De Witt almost four decades ago. We proceed to recapitulate various prescriptions for quantizing gravitation and then discuss simple derivation of the Wheeler-De Witt (WDW) equation in general

relativity and some of its applications in the study of quantum cosmology. There are, broadly speaking, three different approaches to quantize gravity.

The general theory of relativity has been tested to great degree of accuracy in the classical regime. The geometrical description of spacetime plays a cardinal role in Einstein's theory. Therefore, the general relativists emphasize the geometrical attributes of the theory and the central role played by the spacetime structure in their formulation of quantum theory. It is natural to adopt a background-independent approach. In contrast, the path followed by quantum field theorists, where the prescription is valid in the weak-field approximation, the theory is quantized in a given background, usually the Minkowskian space. It is argued by the proponents of the geometric approach, that the background metric should emerge from the theory in a self-consistent manner rather than being introduced by hand when we quantize the theory. One of the earliest attempts to quantize gravity was to follow the route of canonical method. The canonical quantization approach has many advantages. One of the important features is that it is quite similar to the prescriptions adopted in quantum field theory where one uses notion of operators, commutation relations, etc. Moreover, the subtleties encountered in quantizing gravity are transparent. Therefore, the canonical procedure is preferred over the path-integral formulation, although the latter has its own advantages too. Another positive aspect of the canonical approach is that the requirement of background-independent

formulation could be maintained to some extent. Thus, there is room for exploring some of the nonperturbative attributes of the theory. The relativists favor canonical formulation, since some of the geometrical features of general theory of relativity could be incorporated here and be explored to see how far the quantum theory captures such properties of the classical theory. As we shall discuss in sequel, some of the interesting issues of quantum cosmology are addressed in this approach. However, there are limitations and shortcomings in this formulation and we refer the reader to the text books and review articles for further reading and critical assessments of canonical approach to quantize gravity.

The second approach is primarily the endeavor of physicists who have devoted their research to quantum field theory. Feynman's seminal work on quantization of gravity from this perspective has profoundly influenced the subsequent developments. The quantization of gravity is carried out in the weak-field approximation such that the graviton is identified as the fluctuation over the Minkowski background metric. It is a massless spin-2 field as one concludes from the properties of low-energy gravitational interaction in the classical limit. Furthermore, the gauge invariance associated with a spin-2 massless field gets intimately related with invariance of Einstein's theory under general coordinate transformation. In this setup, the field-theoretic techniques could be employed to quantize theory and to consider perturbative expansions for the scattering amplitudes. It is realized that low-energy amplitudes computed from the massless spin-2 theory match with those derived from the Einstein-Hilbert action in the weak-field approximation. Furthermore, the theory is not perturbatively renormalizable since the coupling constant carries dimension. One of the most important outcomes of the investigations from this perspective is the discovery, due to Feynman, that the introduction of ghost fields is necessary in order to maintain unitarity of the S -matrix when one goes beyond the tree level. As is well known, this work has profoundly influenced frontiers of research in physics leading to quantization of Yang-Mills theory which, in turn, paved way for electroweak theory and the QCD. It is worthwhile to mention in passing that the quantum phenomena associated with gravity in the nonperturbative regime cannot be addressed in this framework.

In recent years, superstring theory has been at the center stage in order to provide a unified theory of fundamental interactions. It is postulated that all elementary constituents of matter and the carriers of the interactions such as gauge bosons and graviton are excitations of one-dimensional extended objects:

the strings. The superstring theories are perturbatively consistent in critical ten dimensions. The closed-superstring spectrum contains a spin-2 massless state which is identified to be the graviton. It is well known that perturbative computation of processes involving graviton turn out to be finite. Moreover, the Einstein-Hilbert term appears naturally when one derives the string effective action. Therefore, it is expected that string theory will be able to provide answers to questions related to quantum gravity. Indeed, the theory has met with success in resolving some important issues. We note that cosmological scenario has been discussed in the string theory framework and the WDW equation has played an important role in study of quantum string cosmology. We shall comment on this aspect towards the end of this article.

The Canonical Structure of Einstein Gravity

The Einstein-Hilbert action is

$$S = \frac{1}{16\pi G} \int_M \sqrt{-g} d^4x (R - 2\Lambda) \quad [1]$$

where R is the Ricci scalar derived from the metric, $g_{\mu\nu}$, and Λ is the cosmological constant. The field equations are derived from the action by the standard variational technique. Note that R involves second derivative of the metric. If we have compact manifolds with boundary ∂M such that variations of the metric vanish on the boundary and the normal derivatives do not, it is necessary to add a surface term to this action. The exact form of this term will be discussed later. The Einstein's theory of gravitation is manifestly covariant. The associated action [1] is invariant under general coordinate transformations: under $x^\mu \rightarrow x'^\mu(x)$,

$$g'^{\mu\nu}(x') = g^{\rho\lambda}(x) \frac{\partial x'^\mu}{\partial x^\rho} \frac{\partial x'^\nu}{\partial x^\lambda} \quad [2]$$

Therefore, we expect that the theory will be endowed with constraints expressed in terms of the canonical variables. One can implement general coordinate transformations so that there are only two pairs of canonical phase-space variables on a spacelike hypersurface. In other words, from physical considerations, graviton has only two polarizations whereas the metric has ten components. Therefore, the two physical degrees of freedom can be obtained using the freedom of choosing the "gauge" transformations in this context. It is desirable to identify the constraints and analyze their structure, most appropriately in Dirac's

formalism, and to quantize the theory canonically as the next step. This is the path we intend to follow in order to arrive at the WDW equation.

The Classical Constraints

The Hamiltonian approach is most appropriate to employ the constraint formalism due to Dirac. We recall that the Lagrangian formulation is manifestly covariant as is reflected in the field equations; whereas the spacetime covariance is lost in the passage to the Hamiltonian approach. Furthermore, the spatial components of the metric are the dynamical degrees of freedom. We adopt the formalism introduced by Arnowitt, Deser, and Misner (ADM) for the so-called 3 + 1 split of the hyperbolic Riemannian spacetime metric, $g_{\mu\nu}$. One introduces the lapse function, N^\perp , and the shift function, N^i . We suppress the factors of $1/16\pi G$, etc., for the time being for the general discussions and shall reintroduce them later. The family of spacelike hypersurfaces, Σ_t , are constructed, with metric h_{ij} induced on it. Here t is a timelike parameter, parametrize Σ_t . The distance between points on two neighboring hypersurface, Σ_t and Σ_{t+dt} , with coordinates (t, x^i) and $(t + dt, x^i + dx^i)$, respectively, is given by

$$ds^2 = -(N^\perp)^2 dt^2 + h_{ij}(N^i dt + dx^i)(N^j dt + dx^j) = g_{\mu\nu} dx^\mu dx^\nu \quad [3]$$

The indices of tensors defined on Σ_t are raised and lowered by h_{ij} and its inverse h^{ij} . The relations between the components of $g_{\mu\nu}$ and N^\perp, N^i, h_{ij} can be obtained easily,

$$g_{00} = h_{ij}N^iN^j - (N^\perp)^2, \quad g_{0i} = h_{ij}N^j \quad [4]$$

The above relations can be inverted to give

$$N^i = h^{ij}g_{0j} \quad N^\perp = \frac{1}{\sqrt{-g^{00}}} \quad [5]$$

The relations between spatial components, g_{ij} , of $g_{\mu\nu}$ and h_{ij} and some other useful relations are listed below for later conveniences:

$$\begin{aligned} g^{ij} &= h^{ij} - \frac{N^iN^j}{(N^\perp)^2} \\ \sqrt{-g} &= N^\perp\sqrt{h} \\ g^{0i} &= \frac{N^i}{(N^\perp)^2} \end{aligned} \quad [6]$$

Note that (N^\perp, N^i) are introduced to specify the deformation of the hypersurface and therefore, the evolution equations through the Hamiltonian will not determine them; they are arbitrary functions.

Consequently, [4] implies that g_{00} and g_{0i} will enter the Hamiltonian as arbitrary functions. As alluded to above, h_{ij} and their conjugate momenta π^{ij} are the dynamical degrees of freedom. We may choose $(N^\perp, N^i) = N^\mu$ and h_{ij} as independent variables rather than $(g_{00}, g_{0i}) = g_{0\mu}$ and h_{ij} for convenience and go back to the other set of variables through [4] and [5] if we desire. Let π_μ be canonically conjugate momenta to N^μ , then it is obvious that a Lagrangian multiplier, χ^μ , is necessary so that $\pi_\mu\chi^\mu$ term has to be supplemented to the Hamiltonian due to the arbitrariness of N^μ . We remind the reader that in electrodynamics an analogous situation arises while analyzing its canonical structure – local gauge symmetry plays a crucial role there. It is obvious that the generic form of the Hamiltonian is (we shall introduce $1/16\pi G$, etc., later)

$$H = \int d^3x (N^\perp \mathcal{H}_\perp[h_{ij}, \pi^{ij}] + N^i \mathcal{H}_i[h_{ij}, \pi^{ij}] + \chi \cdot \pi) \quad [7]$$

From the perspective of constraint analysis, it is natural that $\pi^\mu \approx 0$ appears as a first-class constraint as they are multiplied by arbitrary functions. Moreover, this constraint must hold good under the deformation of the surface which implies $\{\pi^\mu, H\}_{PB}$ must vanish weakly leading to $\mathcal{H}_\mu \approx 0$. As a consistency requirement, these must be first-class constraints if N^μ are to be arbitrary functions. We identify that $\pi^\mu \approx 0$ and $\mathcal{H}_\mu \approx 0$ are the primary and secondary constraints, respectively. Thus far, we have discussed the case for pure gravity; the presence of matter fields in the full action modifies the treatment appropriately.

Let us analyze the structure of the constraints for the Einstein-Hilbert action [1]. For a compact manifold with boundary ∂M , we have to add the surface term which takes the form:

$$\frac{1}{8\pi G} \int_{\partial M} d^3x \sqrt{h} K$$

Here K stands for the trace of the extrinsic curvature of the boundary 3-surface and $h = \det h_{ij}$; note that h_{ij} is the induced metric on the 3-surface. If we include matter fields, the corresponding action is to be taken into account. Once we make the 3 + 1 split of the metric, the action assumes the following form:

$$S = \frac{1}{16\pi G} \int d^3x dt N^\perp \sqrt{h} \times (K_{ij}K^{ij} - K^2 + {}^3R - 2\Lambda) \quad [8]$$

where

$$K_{ij} = \frac{1}{N^\perp} \left(-\frac{\partial h_{ij}}{\partial t} + D_i N_j + D_j N_i \right) \quad [9]$$

Here $D_i N_j$ represents covariant derivative of N_j with the connections computed from h_{ij} and 3R is curvature of the 3-surface. The canonical momenta are

$$\pi^{ij} = \frac{\sqrt{h}}{16\pi G} \left(K^{ij} - h^{ij} K^l_l \right) \quad [10]$$

and we can invert this relation to get

$$K^{ij} = -\frac{1}{16\pi G \sqrt{h}} \left(\pi^{ij} - \frac{1}{2} h^{ij} \pi^l_l \right)$$

The Hamiltonian form of action is given by

$$S_H = \int d^3x dt \left(\dot{h}_{ij} \pi^{ij} - N_\perp \mathcal{H}^\perp - N^i \mathcal{H}_i \right) \quad [11]$$

Notice that [8] does not involve time derivatives of N^\perp and N^i , their corresponding canonical momenta vanish.

$$\pi_\perp \approx 0, \quad \pi_i \approx 0 \quad [12]$$

as expected from our earlier discussions about the role of N^μ . A straightforward constraint analysis leads to the pair of constraints

$$\mathcal{H}_i = -2D_j \pi^j_i \approx 0 \quad [13]$$

$$\begin{aligned} \mathcal{H}_\perp = & \frac{16\pi G}{\sqrt{h}} \left(h_{ij} h_{kl} - \frac{1}{2} h_{ik} h_{jl} \right) \pi^{ik} \pi^{jl} \\ & - \frac{\sqrt{h}}{16\pi G} {}^3R \approx 0 \end{aligned} \quad [14]$$

We mention in passing that the above constraint equations get modified in the presence of matter fields in the theory. This is relevant. The WDW equation plays an important role in quantum cosmology to describe the evolution of the universe in early epochs and the equation is studied in the presence of a generic matter content, that is, a scalar field with potential. The constraint equations [13] and [14] modify to

$$\mathcal{H}_i^T = \mathcal{H}_i + \mathcal{H}_i^{\text{matter}} \approx 0 \quad [15]$$

$$\mathcal{H}_\perp^T = \mathcal{H}_\perp + \mathcal{H}^{\text{matter}} \approx 0 \quad [16]$$

The Algebra of Constraints

In order to compute the classical Poisson bracket algebra of the constraints [13] and [14], we use the canonical Poisson bracket relations for the phase-space variables on Σ_t :

$$\{h_{ij}(\mathbf{x}), h_{kl}(\mathbf{x}')\} = 0 \quad [17]$$

$$\{\pi^{ij}(\mathbf{x}), \pi^{kl}(\mathbf{x}')\} = 0 \quad [18]$$

$$\{h_{ij}(\mathbf{x}), \pi^{kl}(\mathbf{x}')\} = \delta^k_i \delta^l_j \delta(\mathbf{x}, \mathbf{x}') \quad [19]$$

Thus, Poisson brackets among the constraints [13] and [14] are

$$\begin{aligned} \{\mathcal{H}_i(\mathbf{x}), \mathcal{H}_j(\mathbf{x}')\} = & -\mathcal{H}_j(\mathbf{x}) \partial_i^x \delta(\mathbf{x}, \mathbf{x}') \\ & + \mathcal{H}_i(\mathbf{x}') \partial_j^x \delta(\mathbf{x}, \mathbf{x}') \end{aligned} \quad [20]$$

$$\{\mathcal{H}_i(\mathbf{x}), \mathcal{H}_\perp(\mathbf{x}')\} = \mathcal{H}_\perp(\mathbf{x}) \partial_i^x \delta(\mathbf{x}, \mathbf{x}') \quad [21]$$

$$\begin{aligned} \{\mathcal{H}_\perp(\mathbf{x}), \mathcal{H}_\perp(\mathbf{x}')\} = & h^{ij}(\mathbf{x}) \mathcal{H}_i(\mathbf{x}) \partial_j^x \delta(\mathbf{x}, \mathbf{x}') \\ & - h^{ij}(\mathbf{x}') \mathcal{H}_i(\mathbf{x}') \partial_j^x \delta(\mathbf{x}, \mathbf{x}') \end{aligned} \quad [22]$$

When we resort to canonical quantization, the starting point is the Hamiltonian action in the first-order formalism, where the canonical variables are subjected to the constraints [13] and [14] in terms of \mathcal{H}_\perp and \mathcal{H}_i satisfying the algebra given by [20]–[22]. One encounters a number of important issues while proceeding to canonically quantize the theory. We shall mention only a few of them in what follows. It is important to address issues related to the role of the constraints in the quantized theory and how to deal with the Lagrange multipliers N^\perp and N^i . A simple proposal is to solve the constraints at the classical level and identify the physical degrees of freedom and quantize the theory subsequently. There are four constraints (first class), $\mathcal{H}_\perp, \mathcal{H}_i$, therefore, out of the 12 phase-space variables, (h_{ij}, π^{ij}) , only eight are independent. We need to supply four gauge conditions in order to render the theory (classically) solvable. Thus, we are left with four physical degrees of freedom in the Hamiltonian phase space and we can quantize them. The implementation of this idea is easier said than done. One obstacle is that the constraints cannot be solved in a closed form in this formalism. If we fix a gauge and quantize the theory, we obviously break the gauge invariance. It is essential to show, subsequently, that all physically observable quantities are independent of the gauge choice. Another criticism of this formalism is that we already get rid of some of the components of the metric. Therefore, the spirit of the general theory of relativity, which is based on the geometrical structure of spacetime, is somewhat diluted. There are other suggestions where h_{ij} and their conjugate momenta are elevated to quantum status before supplying the gauge conditions. The issues of gauge fixing and dealing with the constraints are addressed at the quantum level. We replace the canonical Poisson bracket

algebra by the canonical commutators and proceed further. The momentum operator assumes the form

$$\hat{\pi}^{ij} = -i\hbar \frac{\delta}{\delta h_{ij}}$$

and the wave functional depends on h_{ij} that is, $\Psi[h]$. There are many technical problems related to the properties of the states and we shall not deal with them due to limitations of space. It is essential to discuss the role of the constraints in the quantum theory. We demand that the quantum constraints annihilate the physical states (recall the Gauss law constraint in gauge theories). However, the issue of operator ordering is to be dealt with which in turn is connected with the Hermiticity properties of the quantum constraints. The Hamiltonian constraint $\mathcal{H}_\perp \approx 0$ (henceforth denoted as \mathcal{H} and defined as the Hamiltonian) is a product of the metric h_{ij} and π^{ij} . There is certain ambiguity in defining the constraint. Therefore, one has to choose a convention. The condition that the Hamiltonian, $\hat{\mathcal{H}}^T$, consisting of gravitational and matter components, annihilates the state is expressed as

$$\hat{\mathcal{H}}^T \Psi = 0 \tag{23}$$

When we adopt coordinate representation for π^{ij} , the above equation takes the form

$$\left[-16\pi G G_{ijkl} \frac{\delta}{\delta h_{ij}} \frac{\delta}{\delta h_{kl}} - \frac{\sqrt{h}}{16\pi G} ({}^3R - 2\Lambda) + \mathcal{H}^{\text{matter}} \right] \Psi[h, \phi] = 0 \tag{24}$$

This is the celebrated WDW equation. Here we have considered a simple case where matter Hamiltonian density generically contains a single scalar field, ϕ , and therefore Ψ is functional of 3-metric on Σ_t and ϕ . G_{ijkl} is the De Witt metric in the superspace:

$$G_{ijkl} = \frac{1}{\sqrt{h}} (h_{ik}h_{jl} + h_{il}h_{jk} - h_{ij}h_{kl}) \tag{25}$$

Remarks The space of all 3-metrics and the scalar field (h_{ij}, ϕ) , on Σ_t , for the description of classical evolutions is called the superspace (no connection with the superspace of supersymmetry). Thus, $\Psi[h_{ij}, \phi]$ is a functional on superspace. Furthermore, Ψ carries no explicit dependence on t . This is a consequence of the fact that “time” plays the role of a parameter in the general theory of relativity, thus the dynamical variables h_{ij} and ϕ already provide the evolutionary processes although t does not make its appearance. As mentioned earlier, we always discuss the case when Σ_t is compact. Another point to note

is that the quantum momentum constraint, $\hat{\mathcal{H}}_i$, as an operator annihilates the wave function which is a statement of the quantum-mechanical invariance of the theory under three-dimensional diffeomorphisms. However, the WDW equation conveys invariance of the theory under reparametrization, although careful analysis is necessary to prove this point. Now we proceed to discuss the solutions of the WDW equation.

WDW Equation and the Solutions

It is recognized that the WDW equation [24] is a second-order hyperbolic functional differential equation and naturally it has enormous number of solutions. Therefore, if we want the WDW equation to have any predictive power, it is necessary to introduce boundary conditions. One of the possible choice is to specify the wave function on the boundary of the superspace. Indeed, the central issue of quantum cosmology is about the choice of various boundary conditions which has been an important topic of debates. This point will be briefly discussed later. Notice that the boundary condition has to be introduced keeping in mind how the universe is expected to behave as it evolves. There is a proposition that the boundary condition for the quantum evolution of the universe be given the status of a physical law. Therefore, the role of the wave functional, $\Psi[h_{ij}(x), \phi(x), B]$, its evolution, and interpretation are central to the development of quantum cosmology. Thus, Ψ represents the amplitude for the universe to have $h_{ij}(x)$ on the 3-surface, B , and matter field $\phi(x)$. It is argued that path-integral formalism should be adopted as an alternative to the canonical prescription to solve for the wave function, rather the transition amplitude, satisfying the WDW equation. Here the first step is to define the Euclidean version of the gravitational action keeping in mind the subtleties. As is well known, we deal with propagator (or transition amplitude) in the path-integral approach where the functional integral is carried out over a set of 4-metrics and matter fields with Euclidean action inside the integral acting as the weight factor. We recall that while formulating quantum mechanics in the path-integral approach, we sum over all possible paths in the functional integral. However, in the semiclassical approximation, the amplitude is dominated by the action corresponding to the classical path and we approximate the wave function as $\psi \sim e^{(i/\hbar)S_{cl}}$ and it gets modified appropriately in the Euclidean formulation. In this background, we briefly discuss how the wave function of the universe is obtained in the path-integral formalism.

According to the proposal of Hartle and Hawking, one adopts path-integral formalism for the Euclidean action where the functional integral is not only carried out over the 4-metric, $g_{\mu\nu}$, and the scalar field ϕ , but also one takes sum over the class of manifolds, M . Note that B is a part of the boundary of this set of manifold. If \bar{h}_{ij} and $\bar{\phi}$ are the induced metric and the configuration of the scalar field, ϕ , on the boundary, B , then the propagator (henceforth we just call it the wave function) $\Psi[\bar{h}_{ij}, \bar{\phi}, B]$ can be given a functional-integral representation. Indeed, obtaining the most general form of the path integral, summing over the 4-manifolds, is quite a formidable task. On the other hand, if one chooses a class of 4-manifolds which can be decomposed as a product (foliation) $\mathbf{R} \times B$, the wave function is expressed as

$$\Psi[\bar{h}, \bar{\phi}, B] = \int \mathcal{D}N^\mu \int \mathcal{D}h_{ij} \mathcal{D}\phi f(N^\mu) \Delta_{\text{FPE}}^{-S_E[g_{\mu\nu}, \phi]} \quad [26]$$

We have introduced the gauge-fixing condition as $f(N^\mu)$, which is usually taken to be $N^\mu = l^\mu$ and then the corresponding Faddeev–Popov determinant, Δ_{FP} , has to be inserted into the path-integral measure. We recall from our earlier discussions that N^μ has to be unrestricted on the boundary, B , since they have no dynamical role when we express the action in terms of the variables defined on the 3-surface. As noted in the previous discussion, explicit time dependence does not appear after the $3+1$ split and $(h_{ij}(\mathbf{x}), \phi(\mathbf{x}))$ have no dependence on t . Therefore, we introduce a parameter to designate the paths over which the functional integral is to be taken. Recall that in the quantum-mechanical case, the paths are parametrized as $q_i(t)$ for the coordinates. However, when we resort to a parametrization of the variables for the case at hand, certain conditions must be fulfilled. We are permitted to integrate over h_{ij} and ϕ over only those paths, while parametrizing them as $(h_{ij}(\mathbf{x}, \tau), \phi(\mathbf{x}, \tau))$, so that they match the arguments of the wave function on the boundary B . Therefore, we may define the metric and the scalar field configuration so that at $\tau=1$ they assume their functional values on the boundary: in other words, $\bar{h}_{ij}(\mathbf{x}) = h_{ij}(\mathbf{x}, \tau=1)$ and $\bar{\phi}(\mathbf{x}) = \phi(\mathbf{x}, \tau=1)$. It is worthwhile to go back to the quantum-mechanical analogy once more. When we compute amplitudes/propagators in quantum mechanics, the functional integral is defined for the amplitude of going from a configuration q_i to q_f while summing over all possible paths originating from one endpoint q_i and ending at the final point q_f . On this occasion, we have imposed the constraint on the final endpoint belonging to the

boundary B . Thus, in order to determine the wave function of the universe, we are required to specify the initial configurations of h_{ij} and ϕ at $\tau=0$. We shall not enter into important issues related with the properties of the Euclidean action, the problems associated with the choice of contours of the path integrals, and related topics. The reader will find detailed discussions in the lectures and monographs referred in the “Further reading” section.

It is important to re-emphasize that boundary conditions are to be introduced while solving the WDW equation. It was argued by De Witt that the wave function will be determined uniquely from the mathematical consistency of the theory and that hope has not been realized. Whether one attempts to solve the functional differential WDW equation or obtain the wave function in the path-integral formalism, the issue of boundary condition is unavoidable. There are mainly three different kinds of boundary conditions in quantum cosmology: Hartle–Hawking (HH) no-boundary proposal, Vilenkin’s tunneling mechanism, and Linde’s boundary condition. We shall briefly discuss the first two proposals. Instead of stating the boundary conditions in full generality, we shall envisage quantum cosmology in a minisuperspace and provide illustrative examples to compare the main features of HH and Vilenkin solutions to the WDW equation.

It is realized that the discussion and solutions of quantum cosmology in the superspace is rather difficult, since we deal with functional differential equations and the configuration space is infinite dimensional. Therefore, it is worthwhile to consider a system, as a simple model, which has finite degrees of freedom. Thus, we assume that the metric and matter fields depend only on cosmic time to begin with. There is a physical motivation behind this assumption, since the present classical state of the universe is described by the Friedmann–Robertson–Walker (FRW) metric corresponding to an isotropic and homogeneous universe. Notice that the classical evolution equation resembles that of the motion of a particle. The quantum evolution equations are now given by differential equations of quantum mechanics rather than functional differential equations. Similarly, the path-integral formulation becomes analogous to the quantum-mechanical frame work. Of course, adopting such a simplified approach deprives us from describing some of the important aspects of quantum gravity. However, within this framework, several essential features can be exhibited and deep insight might be gained into the physics of the very early universe. The first step in getting the minisuperspace metric is to assume that the lapse is homogeneous, that is, $N^\perp = N^\perp(t)$

and the shift is set to zero, $N^i = 0$. Thus, the metric takes the form

$$ds^2 = -(N^\perp(t))^2 dt^2 + h_{ij}(\mathbf{x}, t) dx^i dx^j \quad [27]$$

The relevant choice of 3-metric for FRW isotropic and homogeneous universe is

$$h_{ij}(\mathbf{x}, t) dx^i dx^j = a(t)^2 d\Omega_3^2 \quad [28]$$

Note that $d\Omega_3^2$ is the metric on a 3-sphere. It is straightforward to derive the Friedmann equations for such a geometry.

The HH no-boundary condition can be interpreted as a topological proposition about the set of path over which we have to sum. The 3-surface B is to be taken as the only surface of compact 4-manifold M which is endowed with the metric $g_{\mu\nu}$, and \bar{h}_{ij} and $\bar{\phi}$ are the induced metric and the scalar field on the surface. The wave function is obtained by using the matching condition supplemented with initial condition. For the minisuperspace case, initial conditions impose constraints on the scale factor $a(\tau = 0)$ and $(da/d\tau)(\tau = 0)$, and N^\perp is to be gauge fixed. These conditions are to be implemented in the context of determining the wave function of the universe. In the case of the tunneling boundary condition of Vilenkin, the qualitative scenario is as follows. If we look at the solution to the WDW equation (in the path-integral approach, Vilenkin considers Lorentzian action), the solution, crudely speaking, has both ingoing and outgoing modes at the boundary. In his proposal, the outgoing mode at the boundary is to be accepted. The exact prescription is lot more subtle than the above statement, since one has to define the meaning of outgoing mode carefully in the absence of a timelike Killing vector when we write the WDW equation on the superspace. The qualitative picture for Vilenkin's boundary condition, in the minisuperspace, is like tunneling solutions in quantum mechanics when a particle penetrates through a potential barrier.

Let us consider a minisuperspace model with the scalar field and potential $V(\phi)$. The action is

$$S = \frac{1}{2} \int dt a^3 \left[-\frac{1}{N^\perp} \left(\frac{\dot{a}}{a} \right)^2 + \frac{(\dot{\phi})^2}{N^\perp} - N^\perp V(\phi) + \frac{N^\perp}{a^2} \right] \quad [29]$$

A few comments are in order here. For the FRW metric, we have $\sqrt{g}R = 6(-a\ddot{a} + k\dot{a}) +$ a total

derivative term; the total derivative term can be removed by adding a boundary term and k is positive since we take the spatial part to be closed. We have redefined the scale factor, the scalar field, the potential term, and k such that the Einstein-Hilbert action with matter field assumes the form of [29] and this action facilitates the definition of conjugate momenta without cumbersome numerical factors, and the Hamiltonian takes a simple form. The conjugate momenta and resulting Hamiltonian are

$$\pi_a = -\frac{a\dot{a}}{N^\perp}, \quad \pi_\phi = \frac{a^3\dot{\phi}}{N^\perp} \quad [30]$$

$$H_c = \frac{N^\perp}{2} \left[-\frac{\pi_a^2}{a} + \frac{\pi_\phi^2}{a^3} + a^3 V(\phi) - a \right] = N^\perp H \quad [31]$$

and the constraint is $H = 0$. In the quantum cosmology context, we solve the WDW equation: $H\Psi = 0$. Since the exact solution is not possible, one resorts to some approximation with simple assumptions. The differential equation is

$$\left[\frac{\partial^2}{\partial a^2} - \frac{1}{a^2} \frac{\partial^2}{\partial \phi^2} + a^4 V(\phi) - a^2 \right] \Psi = 0 \quad [32]$$

Let us consider the case when $V(\phi)$ does not grow very fast, that is, $V(\phi)/V(\phi)' \ll 1$ and consider the solution to the WDW equation where Ψ has weak dependence on ϕ . Consequently, we may ignore the ϕ derivative term in [32]. The purpose of these assumptions is to reduce the problem to a one-dimensional quantum mechanics problem and then employ WKB method. It is hoped that at least some of the nonperturbative aspects can still be captured. When the effective potential appearing in [32] is negative (this is a classically inaccessible region), the wave function is

$$\Psi(a, \phi) \approx e^{\pm(1/3V(\phi))(1-a^2V(\phi))^{3/2}} \quad [33]$$

We expect the wave function have oscillatory behavior in the classically allowed domain and it does have that property,

$$\Psi(a, \phi) \approx e^{\pm(i/3V(\phi))(a^2V(\phi)-1)^{3/2}} \quad [34]$$

The choice of signs is decided from the boundary conditions imposed and the usual matching of the wave functions of the two regions is done as is the case with the WKB approximation. Note that we are considering the metric and the scalar field on

the boundary which were denoted by \bar{h}_{ij} and $\bar{\phi}$; strictly speaking, we should denote the solutions as \bar{a} and $\bar{\phi}$. But from now on, we drop this bar on a and ϕ .

Let us momentarily assume that V is ϕ -independent and therefore, we have an effective cosmological constant. The problem is identical to the motion of a particle in a potential well. There are two turning points. In one region, the particle starts from $a=0$, reaches one turning point r_1 and returns back. In another case, it starts from $a=\infty$, travels up to $a=r_2$ and reflects back. In the quantum-mechanical case, the particle can tunnel through the barrier. The wave function has both decaying and growing modes under the barrier, and boundary conditions tell us which mode to choose. One possibility is that the particle starts from $a=0$, tunnels through and proceeds towards $a=\infty$, that is, it has outgoing mode. The other possibility is that the wave function has both outgoing and ingoing modes. In this simple scenario, the former corresponds to Vilenkin's tunneling boundary condition, where the universe is created at $a=0$ and it keeps growing. The latter is HH no-boundary proposal where the wave function has both modes and the universe contracts and expands.

Now we discuss the two boundary conditions in the presence of the potential, with the approximations mentioned above. The proposition of Vilenkin amounts to the following conditions on the wave function: the region of the boundary which is nonsingular is ϕ finite and $a=0$. Other than this domain, either a or ϕ diverge on any other region of the boundary; both can diverge in this singular boundary. Notice from the expression for [33] and [34] that the tunneling region corresponds to $a^2V(\phi) < 1$, whereas, the oscillatory domain is $a^2V(\phi) > 1$. If we use the saddle-point approximation, $\Psi \approx e^{\pm iS_{cl}}$. Vilenkin's boundary condition corresponds to $\Psi \approx e^{-iS_{cl}}$, with

$$S_{cl} = \frac{(\sqrt{a^2V(\phi)} - 1)^3}{3V(\phi)}$$

So far, we considered the situation where differential operator for ϕ is dropped in [32]. In order to account for weak ϕ -dependence, we could introduce it by multiplying a slowly varying function, say $F(\phi)$ and write $\Psi(a, \phi) \sim F(\phi)e^{-iS_{cl}}$. Similarly, the wave function can be obtained under the barrier and required to satisfy WKB matching conditions. Furthermore, the regularity condition on the wave function in small scale factor limit and behavior of its derivative with respect to ϕ in that limit determine the form of $F(\phi)$. In summary, the

Vilenkin boundary conditions yield the following wave functions:

$$\Psi(a, \phi)_V \approx e^{-(1/3V(\phi))(1-[1-a^2V(\phi)]^{3/2})} \quad [35]$$

$$\Psi(a, \phi)_V \approx e^{-1/3V(\phi)} e^{-(i/3V(\phi))[a^2V(\phi)-1]^{3/2}} \quad [36]$$

Note that [35] is the wave function under the barrier, that is, $a^2V(\phi) < 1$ in this region, whereas [36] is in the classically accessible domain ($a^2V(\phi) > 1$) which is reflected by the oscillatory character. The slowly varying function $F(\phi) \sim e^{-1/3V(\phi)}$ appears as the common factor for the wave functions in the two domains.

The HH no-boundary proposal to derive the wave function of the universe was formulated in the Euclidean path-integral formalism. A considerable amount of attention has been focused in this area. We shall present the HH wave function providing only a sketchy argument. In the Euclidean description, 4-metric is $ds^2 = (N^\perp)^2 d\tau^2 + a^2(\tau) d\Omega_3^2$. The 4-geometry should close in a regular way. If we make the bounding 3-space smaller and smaller, it can be closed with flat space. We can infer about the behavior of the scale factor in the limit $\tau \rightarrow 0$ from this consideration. Furthermore, in the semiclassical approximation $\Psi(a, \phi) \sim e^{-S_E}$; we have replaced $(\bar{a}, \bar{\phi})$ by (a, ϕ) as remarked earlier. Thus, our aim is to evaluate S_E at the saddle point. This is achieved by writing down the (Euclidean version) field equations for a and ϕ and the Hamiltonian constraint, and then solve for $a(\tau)$, $\phi(\tau)$, and $N^\perp(\tau)$. Eventually, we want to eliminate N^\perp and then obtain S_E . After all, the path integral is dominated by the classical trajectory, $a(\tau)$, and one does not fix the gauge for N^\perp while solving for a . In fact, the lapse gets eliminated by utilizing the Hamiltonian constraint which involve τ -derivatives of both a and ϕ . We mention, without going into details, that the classical action is not unique. One of the ways to visualize it is to note that the solutions obtained for the lapse from the Hamiltonian constraint have sign ambiguities.

The classical action is

$$S_E^\pm = -\frac{1}{3V(\phi)} \left(1 \pm [1 - a^2V(\phi)]^{3/2} \right) \quad [37]$$

Note that the two solutions correspond to 3-sphere boundary being closed off by sections of 4-sphere. Moreover, the Euclidean action is negative. Hartle and Hawking argue that the negative sign in [37] gives the correct answer since the wave function peaks for that choice. However, there is no unanimity

for HH argument and some authors have put forward a point of view that additional inputs are necessary to arrive at the HH conclusion about choosing the negative sign for S_E in [37]. We refer the reader to the reviews of Hartle and Halliwell for detailed discussions on the choice of contours for path integrals, subtleties involved in getting various solutions for the lapse and their interpretations. We give below the wave function under the barrier (with choice of negative sign in [37]):

$$\Psi_{HH}(a, \phi) \approx e^{(1/3V(\phi))(1-[1-a^2V(\phi)]^{3/2})} \quad [38]$$

$$\Psi_{HH}(a, \phi) \approx e^{1/3V(\phi)} \times \cos\left(\frac{1}{3V(\phi)}[a^2V(\phi) - 1]^{3/2} - \frac{\pi}{4}\right) \quad [39]$$

Remarks The wave function in [38] is obtained in the classically inaccessible region under the condition $a^2V(\phi) < 1$, and wave function [39] corresponds to the case $a^2V(\phi) > 1$, where the particle motion is permissible classically. Note the factor $e^{1/3V(\phi)}$ in the wave functions in both the regions and compare that with the Vilenkin's wave function which has the opposite sign. We may conclude where the wave function will peak for each of the two boundary conditions. Whereas Vilenkin's proposal implies that $\Psi_V(a, \phi)$ peaks when $V(\phi)$ takes large values, HH no-boundary condition tells us that it peaks when $V(\phi) \rightarrow 0$. Furthermore, we note that Ψ_V is complex and Ψ_{HH} is real in the oscillatory region. Although the debates on the merits and demerits of each of the boundary proposals are going on for more than two decades, the issue is far from being settled. In the absence of any experimental tests, there is no way to favor one boundary proposal over another. Then, boundary conditions do have predictions about the evolution of the universe after the quantum era and have predictions in that (classical) regime. Therefore, determination of the wave function with specific boundary conditions does have some connections with the laws that govern the evolution of our universe in the present epoch.

It is worthwhile to dwell on the WDW equation from the perspectives of string theories. Indeed, there have been important developments to understand the dynamics of the universe in the string-theoretic framework. It is important to note the key role played by dilaton in string theory: (1) it is one of the massless states of the theory, and (2) the vacuum expectation value (VEV) of this field determines the coupling constants we hope to use in describing fundamental interactions. Therefore, the graviton is

always accompanied by the dilaton in any string-theoretic approach to study the universe. The duality symmetries are recognized to provide deep understanding of the string dynamics. Therefore, the investigations of quantum gravity phenomena from the string-theory viewpoint are necessarily influenced by above mentioned facts. Indeed, classical cosmological solutions, derived from string effective action, have several interesting characteristics. We mention is passing that the WDW equation has played an important role to study quantum evolution equations in string cosmology. The choice of operator-ordering prescription in defining the WDW Laplace-Beltrami operator can be resolved by appealing to the duality symmetries. Furthermore, the boundary conditions imposed on the wave function are dictated by string symmetries and therefore, the resulting wave function has very interesting properties. The string theory has addressed some of the most important problems in quantum gravity and it has provided resolutions to several key issues. It is expected that string theory will provide answers to challenging questions in quantum cosmology. In summary, we have conveyed some of the salient aspects of the WDW equation. The canonical quantization technique is adopted to study quantum gravity in this approach. We have illustrated the crucial role of the constraint formalism due to Dirac and argued that some of the nonperturbative aspects of quantum gravity could be retained. In a short article of this nature, it is not possible to provide detailed discussion about the general derivation of the WDW equation and discuss the role of boundary conditions more exhaustively. Instead, we presented some of the key steps in the derivation of the WDW equation adopting the canonical formalism and provided simple examples. The subject is still an active area of research. The interested reader may benefit from the bibliography.

See also: Canonical General Relativity; Loop Quantum Gravity; Quantum Cosmology; Quantum Dynamics in Loop Quantum Gravity; Quantum Geometry and its Applications; Superstring Theories.

Further Reading

- De Witt BS (1967) Quantum theory of gravity. *Physical Review* 160: 1113.
- Feynman RP, Morinigo FB, and Wagner WG (1995) *Feynman Lectures on Gravity*. New York: Addison-Wesley.
- Halliwell JJ (1990) Quantum cosmology. In: Randjbar-Daemi S, Sezgin E, and Shafi Q (eds.) *Summer School in High Energy Physics and Cosmology*, p. 513. Singapore: World Scientific.
- Hartle JB (1989) Introductory lectures on quantum cosmology. In: Coleman S, Hartle JB, Piran T, and Weinberg S (eds.) *Quantum Cosmology and Baby Universes*, Proceedings of 7th Winter School in Theoretical Physics, p. 159. Jerusalem: World Scientific.

Hartle JB and Hawking SW (1983) Wave function of the universe. *Physical Review D* 28: 2960.

Hawking SW (1983) Quantum cosmology. Les Houches Lectures on Quantum Cosmology, p. 333. (Les Houches Publication) in Einstein Centenary Volume.

Vilenkin A (1984) Quantum creation of universe. *Physical Review D* 30: 509.

Wheeler JA (1963) Geometrodynamics. In: De Witt C and De Witt BS (eds.) *Relativity, Groups and Topology*. New York: Gordon and Breach.

Wightman Axioms see Axiomatic Quantum Field Theory

Wulff Droplets

S Shlosman, Université de Marseille, Marseille, France

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Historically, the first question where the Wulff shapes have appeared is the one of the formation of a droplet or a crystal of one substance inside another. The natural problem here is: what shape such a formation would take? The statement that such a shape should be defined by the minimum of the overall surface energy subject to the volume constraint is physically very natural. In the isotropic case, when the surface tension does not depend on the orientation of the surface, and so is just a positive number, the shape in question should be of course spherical (provided we neglect the gravitational effects). In a more general situation the shape in question is less symmetric. The corresponding variational problem is called the Wulff problem. Wulff (1901) formulated it in his paper, where he also presented a geometric solution to it, called the “Wulff construction.”

The Wulff variational problem is formulated as follows. Let $\tau(\mathbf{n})$, $\mathbf{n} \in S^{d-1}$, be some continuous function on the unit sphere $S^{d-1} \subset \mathbb{R}^d$. We suppose that $\tau > 0$, and that τ is even: $\tau(\mathbf{n}) = \tau(-\mathbf{n})$. The value $\tau(\mathbf{n})$ plays the role of the surface tension between two phases separated by the hyperplane orthogonal to the vector \mathbf{n} . For every closed compact (hyper)surface $M^{d-1} \subset \mathbb{R}^d$, we define its surface energy as

$$\mathcal{W}_\tau(M) = \int_M \tau(\mathbf{n}_s) \, ds$$

where \mathbf{n}_s is the normal vector to M at $s \in M$. The functional $\mathcal{W}_\tau(M)$ has the meaning of the surface energy of the M -shaped droplet made from one of these two phases. It is called the Wulff functional. Let \mathfrak{W}_τ be the surface which minimizes $\mathcal{W}_\tau(\cdot)$ over all the surfaces enclosing the unit volume. Such a

minimizer does exist and is unique up to translation. It is called the Wulff shape.

The following is the geometric construction of \mathfrak{W}_τ . Consider the set

$$K_\tau = \left\{ \mathbf{x} \in \mathbb{R}^d: \forall \mathbf{n} \in S^{d-1} (\mathbf{x}, \mathbf{n}) \leq \tau(\mathbf{n}) \right\}$$

If we define the half-spaces

$$L_{\tau, \mathbf{n}} = \left\{ \mathbf{x} \in \mathbb{R}^d: (\mathbf{x}, \mathbf{n}) \leq \tau(\mathbf{n}) \right\}$$

then

$$K_\tau = \bigcap_{\mathbf{n}} L_{\tau, \mathbf{n}} \tag{1}$$

In particular, K_τ is convex. It turns out that

$$\mathfrak{W}_\tau = \lambda_\tau \partial(K_\tau)$$

where the dilatation factor λ_τ is defined by the normalization: $\text{vol}(\lambda_\tau K_\tau) = 1$. The relation [1] is called the Wulff construction. For the future use, we introduce the notation w_τ for the value of the surface energy of the Wulff shape:

$$w_\tau = \mathcal{W}_\tau(\mathfrak{W}_\tau)$$

The Wulff construction was considered by the rigorous statistical mechanics as just a phenomenological statement, though the notion of the surface tension was among its central notions. The situation changed after the appearance of the book by Dobrushin *et al.* (1992). There it was shown that in the setting of the canonical ensemble formalism, in the regime of the first-order phase transition, the (random) shape occupied by one of the phases has asymptotically (in the thermodynamic limit) a nonrandom shape, given precisely by the Wulff construction! In other words, a typical macroscopic random droplet looks very close to the Wulff shape.

In what follows we will explain the above result. Another important application of the concepts introduced above – the role played by the Wulff

shapes in the theory of metastability – is also described (see Metastable States).

Crystals in the Ising Model

Ising spins σ_x take values ± 1 , with $x \in \mathbb{Z}^d$. We will wrap \mathbb{Z}^d into a torus T_N^d by taking a factor lattice: $T_N^d = \mathbb{Z}^d / N\mathbb{Z}^d$. Ising-model grand canonical Gibbs state in T_N^d is the probability measure μ_N^β :

$$\mu_N^\beta(\sigma) = Z_{N,\beta}^{-1} \exp(-\beta H_N(\sigma))$$

Here $H_N(\sigma) = -\sum_{x,y \text{ n.n.}, x,y \in T_N^d} \sigma_x \sigma_y$, $\beta > 0$ is the inverse temperature, and $Z_{N,\beta}$ is the normalization factor. Ising-model canonical Gibbs state in T_N^d is the probability measure $\mu_N^{\beta,\rho}$, obtained from μ_N^β by taking its conditional distribution:

$$\mu_N^{\beta,\rho}(\sigma) = \mu_N^\beta\left(\sigma \mid \sum_{x \in T_N^d} \sigma_x = \rho N^d\right), \quad |\rho| < 1$$

(Here we make a slight abuse of notation. More precisely, since $\sigma_x = \pm 1$, one has to consider the conditioning $\sum \sigma_x = \rho_N N^d$, where $\rho_N \rightarrow \rho$ as $N \rightarrow \infty$, while the numbers $(1 - \rho_N)N^d$ are even integers; otherwise the condition is empty.) We will characterize the canonical state $\mu_N^{\beta,\rho}$ by describing the properties of contours, $\{\gamma_i(\sigma)\}$, of configuration σ . Contours γ_i of configuration σ are hypersurfaces made of elementary $(d-1)$ -dimensional unit cubes of the dual lattice, which separate the nearest-neighbor (n.n.) points $x, y \in T_N^d$ where $\sigma_x \neq \sigma_y$.

Suppose that the temperature β^{-1} is low enough, while the density parameter ρ satisfies the constraints:

$$m_d^*(\beta) > \rho > g_d$$

Here $m_d^*(\beta)$ is the spontaneous magnetization of the d -dimensional Ising model, while g_d is some geometric factor, the role of which will be explained later. The above constraint forces some amount of the $(-)$ -phase into the $(+)$ -phase. It turns out that this amount gathers into one big droplet, which has approximately the Wulff shape.

We first formulate the known rigorous results for the case $d=2$, and then indicate some extensions.

Two-Dimensional Case

The following holds with $\mu_N^{\beta,\rho}$ -probability approaching 1 as $N \rightarrow \infty$:

- The set $\{\gamma_i(\sigma)\}$ of contours of σ has precisely one “big” contour, $\Gamma(\sigma)$; the diameters of other contours do not exceed $K \ln N$, $K = K(\beta)$.
- The area $|\text{Int } \Gamma(\sigma)|$ inside $\Gamma(\sigma)$ satisfies

$$||\text{Int } \Gamma(\sigma)| - \nu N^2| \leq KN^{6/5}(\ln N)^*$$

where

$$\nu = \frac{m_2^*(\beta) - |\rho|}{2m_2^*(\beta)}, \quad \varkappa = \varkappa(\rho)$$

- There is a point $x = x(\sigma)$ – the “center” of $\Gamma(\sigma)$ – such that the shift of $\Gamma(\sigma)$ by $-x(\sigma)$ brings the contour $\Gamma(\sigma)$ very close to the scaled Wulff curve, defined by the Ising-model surface tension τ :

$$r_H\left(\Gamma(\sigma) - x(\sigma), \sqrt{\frac{2\nu}{w_\tau}} NW_\tau\right) \leq KN^{2/3}(\ln N)^* \quad [2]$$

(Here r_H is the Hausdorff distance: for every two sets $A, C \in \mathbb{R}^d$, $r_H(A, C)$ is defined as $\max\{\inf[r: A \subset C + B_r], \inf[r: C \subset A + B_r]\}$, where B_r is the ball of radius r .)

The proof of the above result is the content of the book by Dobrushin *et al.* (1992). In the two-dimensional case, it remains true for all temperatures β^{-1} below the critical one (Ioffe and Schonmann 1998). The value $2/3$ of the exponent is an improvement of the original $3/4$ result (Alexander 1992). Probably, it can be further improved down to $1/2$. Though Dobrushin *et al.* (1992) treat only the Ising model, their results are valid for a wide range of other models.

The restriction $\rho > g_d$ in the theorem is needed because without it the droplet may prefer to assume the shape of a strip between two meridians rather than to take the Wulff shape.

Three-Dimensional Case

In the case $d=3$ or $d \geq 3$, the statement that a typical configuration σ has only one big contour $\Gamma(\sigma)$ is still true. But the analog of [2] is not known. It is natural to conjecture that it holds at low temperatures, even in a stronger version, with only a logarithmic term $K(\ln N)^*$ in the RHS. It probably fails at higher subcritical temperatures.

What is known to hold is a weaker version of this theorem, where the distance between random droplet and the Wulff shape is measured not in Hausdorff distance, but in L^1 sense. To state the corresponding theorem, we will associate with every configuration σ on a lattice torus T_N^d a real-valued function $M_\sigma(t)$ on the unit torus $\mathbb{T}^d = \mathbb{R}^d / \mathbb{Z}^d$, and we then compare this function with the indicator function \mathbb{I}_{sK_τ} , where $sK_\tau \subset \mathbb{T}^d$ is the Wulff body, properly scaled.

The function $M_\sigma(t)$ is defined as follows. We denote by i_N the natural embedding of the discrete torus T_N^d into \mathbb{T}^d , the image of i_N being the grid with spacing $1/N$. For $t \in \mathbb{T}^d$ we define $b_N(t) \subset \mathbb{T}^d$ to be

the ball centered at t with radius $\sqrt[d]{1/N}$, and let $B_N(t) \subset \Lambda(N)$ be its preimage under i_N . Then

$$M_\sigma(r) = \frac{1}{|B_N(t)|} \sum_{x \in B_N(t)} \sigma(x)$$

We have to expect to see a droplet sK_τ with

$$s = \sqrt[d]{\frac{d}{w_\tau} \frac{m_d^*(\beta) - \rho}{2m_d^*(\beta)}}$$

Let us introduce for every subset $A \subset \mathbb{T}^d$ the indicator

$$I_A(t) = \begin{cases} 1, & t \in A \\ -1, & t \in A^c \end{cases}$$

For every function ν in $L^1(\mathbb{T}^d)$ we denote by $U(\nu, \delta)$ its δ -neighborhood in $L^1(\mathbb{T}^d)$.

The result can now be formulated. Suppose the temperature β^{-1} is below the critical one. Then the function $M_\sigma(t)$ is close to the characteristic function of the Wulff shape: For every $\delta > 0$

$$\lim_{N \rightarrow \infty} \mu_N^{\beta, \rho} \left\{ \frac{1}{m_d^*(\beta)} M_\sigma(\cdot) \in \bigcup_{t \in \mathbb{T}^d} U(I_{sK_\tau+t}, \delta) \right\} = 1$$

The shifts by all t -s of the Wulff shape sK_τ appear in the statement since the location of the droplet can be arbitrary. Note that if a point t is such that the ball $B_N(t)$ stays away from the boundary of the

droplet $\Gamma(\sigma)$ present in the configuration σ , then the value $M_\sigma(t)$ should be expected to be $\pm m_d^*(\beta)$, depending on whether t is outside or inside the droplet, which explains the factor $1/m_d^*(\beta)$.

For a proof, see Bodineau (1999) and Cerf and Pizstora (1999).

See also: Cluster Expansion; Large Deviations in Equilibrium Statistical Mechanics; Metastable States; Percolation Theory; Statistical Mechanics of Interfaces.

Further Reading

Alexander K (1992) Stability of the Wulff minimum and fluctuations in shape for large clusters in two-dimensional percolation. *Probability Theory and Related Fields* 91: 507–532.

Bodineau T (1999) The Wulff construction in three and more dimensions. *Communications in Mathematical Physics* 207(1): 197–229.

Cerf R and Pizstora A (2000) On the Wulff crystal in the Ising model. *Annals of Probability* 28(3): 947–1017.

Dobrushin RL, Kotecky R, and Shlosman SB (1992) *Wulff Construction: A Global Shape from Local Interaction*. AMS Translations Series. Providence, RI: American Mathematical Society.

Ioffe D and Schonmann R (1998) Dobrushin–Kotecky–Shlosman theory up to the critical temperature. *Communications in Mathematical Physics* 199: 117–167.

Wulff G (1901) Zur frage der geschwindigkeit des wachstums unter auflösung der kristallflächen. *Z. Kristallogr.* 34: 449–530.

Y

Yang–Baxter Equations

J H H Perk and H Au-Yang, Oklahoma State University, Stillwater, OK, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

The term Yang–Baxter equations (YBEs) was coined by Faddeev in the late 1970s to denote a principle of integrability, that is, exact solvability, in a wide variety of fields in physics and mathematics. Since then it has become a common name for several classes of local equivalence transformations in statistical mechanics, quantum field theory, differential equations, knot theory, quantum groups, and other disciplines. We shall cover the various versions and their relationships, paying attention also to the early historical development.

Electric Networks

The first such transformation came up as early as 1899 when the Brooklyn engineer Kennelly published a short paper, entitled “The equivalence of triangles and three-pointed stars in conducting networks.” This work gave the definite answer to such questions as whether it is better to have the three coils in a dynamo – or three resistors in a network – arranged as a star or as a triangle, see Figure 1. Using Kirchhoff’s laws, the two situations in Figure 1 can be shown to be equivalent provided

$$\begin{aligned} Z_1 \bar{Z}_1 &= Z_2 \bar{Z}_2 = Z_3 \bar{Z}_3 \\ &= Z_1 Z_2 + Z_2 Z_3 + Z_3 Z_1 \end{aligned} \quad [1]$$

$$= \bar{Z}_1 \bar{Z}_2 \bar{Z}_3 / (\bar{Z}_1 + \bar{Z}_2 + \bar{Z}_3) \quad [2]$$

Here one has to take either [1] or [2] as second line of the equation, depending on which direction the transformation is to go. The star–triangle transformation thus defined is also known under other names within the electric network theory literature as wye–delta ($Y - \Delta$), epsilon–delta ($\Upsilon - \Delta$), or tau–pi ($T - \Pi$) transformation.

Spin Models

When Onsager wrote his monumental paper on the Ising model published in 1944, he made a brief remark on an obvious star–triangle transformation relating the model on the honeycomb lattice with the one on the triangular lattice. His details on this were first presented in Wannier’s review article of 1945. However, the star–triangle transformation played a much more crucial role in Onsager’s reasoning, as it is also intimately connected with his elliptic function uniformizing parametrization.

Furthermore, it implies the commutation of transfer matrices and spin-chain Hamiltonians. Only in his Battelle lecture of 1970 did Onsager explain how he used this remarkable observation in his derivation of the formula for the spontaneous magnetization which he had announced as a conference remark in 1948 and of which the first complete derivation had been published by Yang in 1952 using a completely different method.

Many other applications and generalizations have since appeared. Most generally, we can consider a system whose state variables – also called spins – take values from some suitable discrete or continuous sets. The interactions between spins a and b are given in terms of weight factors W_{ab} and \bar{W}_{ab} , which are complex numbers in general, see Figure 2. One quantity of special interest is the partition function – sum of the product of all weight factors over all allowed spin values. The integrability of the model is expressed by the existence of spectral variables – rapidities p, q, r, \dots – that live on oriented lines, two of which cross between a and b as indicated by the dotted lines in Figure 2. Arrows from a to b are added to keep track of the ordering of a and b in case the weights are chiral (not symmetric).

In Onsager’s special choice of the Ising model the spins take values $a, b, c, \dots = \pm 1$ and the weight factors are the usual real positive Boltzmann weights depending on the product $ab = \pm 1$, uniformizing variable $p - q$, and elliptic modulus k . In the integrable chiral Potts model the weights depend on $a - b \bmod N$, with $a, b = 1, \dots, N$, whereas the rapidities p and q are living in general on a higher-genus curve.

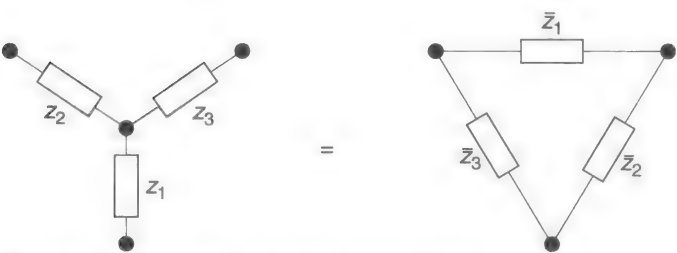


Figure 1 Star-triangle equation for impedances.

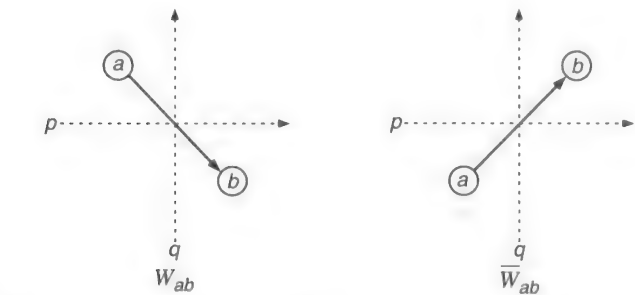


Figure 2 Spin model weights $W_{ab}(p, q)$ and $\overline{W}_{ab}(p, q)$.

When the weights are asymmetric in the spins, there are two sets of star-triangle equations which can be expressed both pictorially (Figure 3) and algebraically:

$$\sum_d \overline{W}_{cd}(p, q) \overline{W}_{db}(q, r) W_{da}(p, r) = R(p, q, r) W_{ba}(p, q) W_{ca}(q, r) \overline{W}_{cb}(p, r) \quad [3]$$

$$\overline{R}(p, q, r) W_{ab}(p, q) W_{ac}(q, r) \overline{W}_{bc}(p, r) = \sum_d \overline{W}_{dc}(p, q) \overline{W}_{bd}(q, r) W_{ad}(p, r) \quad [4]$$

Note that eqns [3] and [4] differ from each other by the transposition of both spin variables in all six weight

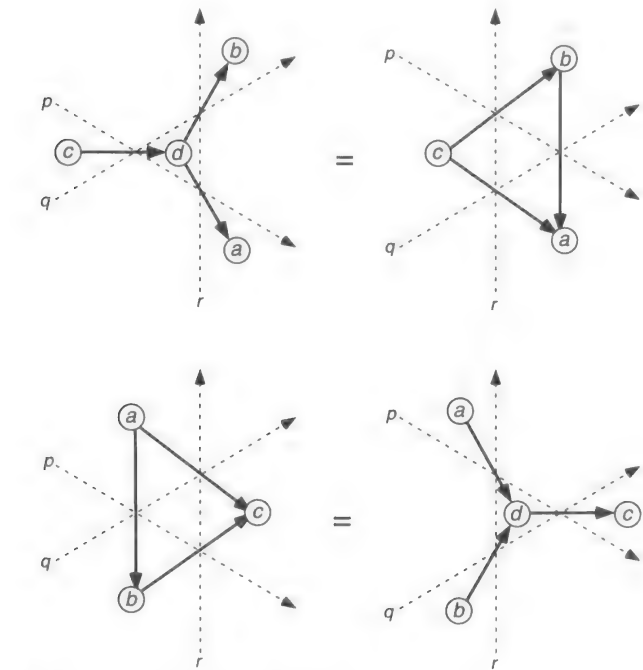


Figure 3 Star-triangle equation.

factors. In general, there may also appear scalar factors $R(p, q, r)$ and $\overline{R}(p, q, r)$, which can often be eliminated by a suitable renormalization of the weights. If a, b , and c take values in the same set, we can sum over $a=b=c$, showing that $R = \overline{R}$ in that case.

The Kennelly star-triangle equation [1], [2] can be recovered as a special limit of a spin model where the states are continuous variables.

Knot Theory and Braid Group

A seemingly totally different situation occurs in the theory of knots, links, tangles, and braids. In 1926, Reidemeister showed that only three types of moves suffice to show the equivalence between two different configurations, see Figure 4. Moves of type I – removing simple loops – do not apply to braids. Moves of type II, for which one strand crosses twice over another strand, can be reformulated for braids, namely that an overcrossing is the inverse of an undercrossing. The Reidemeister move of type III is a precursor of the more general Yang-Baxter moves and can be represented also by the defining relations of Artin’s braid group. Let $R_{i,i+1}$ be the operator representing the situation in which the strand in position i crosses over the one in position $i + 1$. Then a braid can be represented by a product of $R_{j,j+1}$ ’s and their inverses, provided

$$R_{i,i+1} R_{i+1,i+2} R_{i,i+1} = R_{i+1,i+2} R_{i,i+1} R_{i+1,i+2} \quad [5]$$

and

$$[R_{i,i+1}, R_{j,j+1}] = 0, \quad \text{if } |i - j| \geq 2 \quad [6]$$

and similar relations in which $R_{i,i+1}$ and/or $R_{i+1,i+2}$ are replaced by their inverses.

Factorizable S-Matrices and Bethe Ansatz

In the early 1960s, Lieb and Liniger solved the one-dimensional Bose gas with delta-function interaction using the Bethe ansatz. Yang and McGuire then tried to generalize this result to systems with internal degrees of freedom and to fermions. This led to the

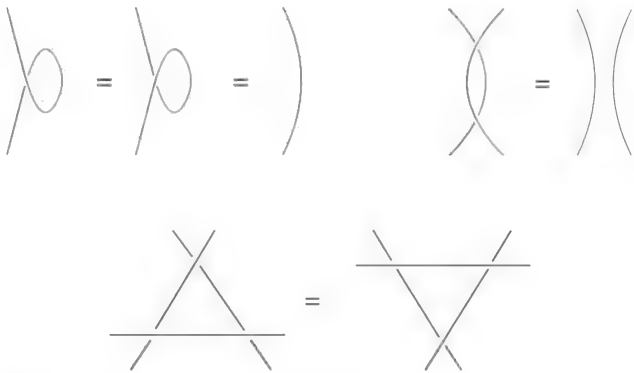


Figure 4 Reidemeister moves of types I, II, and III.

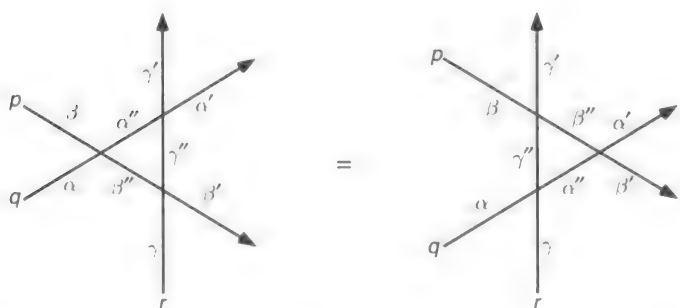


Figure 5 Vertex model YBE.

discovery of the condition for factorizable S -matrices by McGuire in 1964, represented pictorially by Figure 5, where the world lines of the particles are given. Upon collisions the particles can only exchange their rapidities p, q, r , so that there is no dispersion. Also indicated are the internal degrees of freedom in Greek letters. In other words, the three-body S -matrix can be factorized in terms of two-body contributions and the order of the collisions does not affect the final outcome. McGuire also realized that this condition is all one needs for the consistency of factoring the n -body S -matrix in terms of two-body S -matrices. The consistency condition is obviously related to the Reidemeister move of type III in Figure 4.

Yang succeeded in solving the spin-1/2 fermionic model using a nested Bethe ansatz, utilizing a generalization of Artin’s braid relations [5] and [6],

$$\begin{aligned} &\check{R}_{i,i+1}(p-q)\check{R}_{i+1,i+2}(p-r)\check{R}_{i,i+1}(q-r) \\ &= \check{R}_{i+1,i+2}(q-r)\check{R}_{i,i+1}(p-r)\check{R}_{i+1,i+2}(p-q) \end{aligned} \quad [7]$$

He submitted his findings in two short papers in 1967. The \check{R} operators in eqn [7] – a notation introduced later by the Leningrad school – depend on differences of two momenta or two relativistic rapidities. Sutherland solved the general spin case using repeated nested Bethe ansätze, while Lieb and Wu used Yang’s work to solve the one-dimensional Hubbard model.

Vertex Models

Since Lieb’s solution of the ice model by a Bethe ansatz, there have been many developments on vertex models, in which the state variables live on line segments and weight factors $\omega_{\alpha\mu}^{\lambda\beta}$ are assigned to a vertex where four line segments with the four states $\alpha, \mu, \lambda, \beta$ on them meet, see Figure 6.

Baxter solved the eight-vertex model in 1971, using a method based on commuting transfer matrices, starting from a solution of what he then called the generalized star–triangle equation, but what is now commonly called the Yang–Baxter equation (YBE):

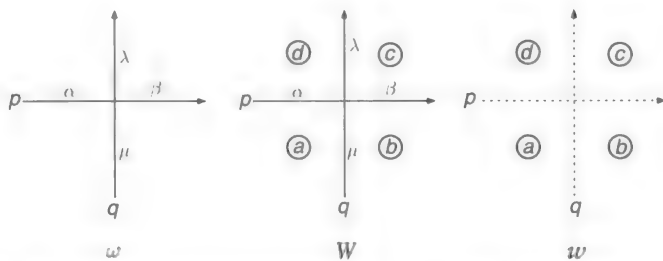


Figure 6 Vertex model weight $\omega_{\alpha\mu}^{\lambda\beta}(p, q)$, mixed model weight $W_{\alpha\mu|\lambda\beta}^{dc}(p, q)$ and IRF model weight $w_{ab}^{dc}(p, q)$.

$$\begin{aligned} &\sum_{\alpha''} \sum_{\beta''} \sum_{\gamma''} \omega_{\beta\alpha}^{\alpha''\beta''}(p, q) \omega_{\alpha''\gamma''}^{\gamma'\alpha'}(q, r) \omega_{\beta''\gamma}^{\gamma''\beta'}(p, r) \\ &= \sum_{\alpha''} \sum_{\beta''} \sum_{\gamma''} \omega_{\beta''\alpha''}^{\alpha'\beta'}(p, q) \omega_{\alpha\gamma}^{\gamma''\alpha''}(q, r) \omega_{\beta\gamma''}^{\gamma'\beta''}(p, r) \end{aligned} \quad [8]$$

This equation is represented graphically in Figure 5. From it one can also derive a sufficient condition for the commutation of transfer matrices and spin-chain Hamiltonians, generalizing the work of McCoy and Wu, who had earlier initiated the search by showing that the general six-vertex model transfer matrix commutes with a Heisenberg spin-chain Hamiltonian. To be more precise, Baxter found that if $\omega_{\alpha\mu}^{\lambda\beta} = \delta_{\alpha}^{\lambda} \delta_{\mu}^{\beta}$ for some choice of p and q , some spin-chain Hamiltonians could be derived as logarithmic derivatives of the transfer matrix.

Interaction-Round-a-Face Model

Baxter introduced another language, namely that of the IRF or “interaction-round-a-face” model, which he introduced in connection with his solution of the hard-hexagon model. This formulation is convenient when studying one-point functions using the corner-transfer-matrix method. Now the integrability condition can be represented graphically as in Figure 7 or algebraically as

$$\begin{aligned} &\sum_d w_{cb'}^{a'd}(p, q) w_{dc'}^{a'b}(q, r) w_{b'a}^{d'c}(p, r) \\ &= \sum_{d'} w_{d'a}^{b'c'}(p, q) w_{b'a}^{c'd'}(q, r) w_{cd'}^{a'b}(p, r) \end{aligned} \quad [9]$$

The spins live on faces enclosed by rapidity lines and the weights $w_{ab}^{dc}(p, q)$ are assigned as in Figure 6.

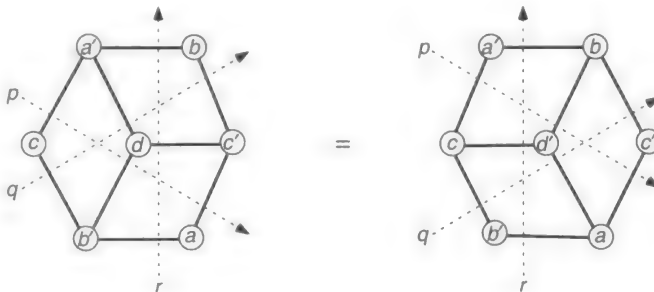


Figure 7 IRF model YBE.

Baxter discovered a new principle based on eqns [8] and [9], which he called Z-invariance, as it expresses an invariance of the partition function Z under moves of rapidity lines. This also implies that typical one-point functions are independent of the values of the rapidities, while two-point functions can only depend on the values of the rapidities of rapidity lines crossing between the two spins considered. Many recent results on correlation functions in integrable models depend on this observation of Baxter.

IRF-Vertex Model

In Figure 6, we have also defined mixed IRF-vertex model weights $W_{\alpha\mu|ab}^{\lambda\beta|dc}(p,q)$. (We could put further state variables on the vertices, but then the natural thing to do is to introduce new effective weights summing over the states at each vertex.) With the choice made a more general YBE can be represented as in Figure 8, or by

$$\begin{aligned} &\sum_{\alpha''} \sum_{\beta''} \sum_{\gamma''} \sum_d W_{\beta\alpha}^{\alpha''\beta''} |_{cb'}^{a'd}(p,q) \\ &\quad \times W_{\alpha''\gamma''}^{\gamma'\alpha'} |_{dc'}^{a'b}(q,r) W_{\beta''\gamma''}^{\gamma''\beta'} |_{b'a}^{d'c'}(p,r) \\ &= \sum_{\alpha''} \sum_{\beta''} \sum_{\gamma''} \sum_{d'} W_{\beta''\alpha''}^{\alpha'\beta'} |_{d'a}^{b'c'}(p,q) \\ &\quad \times W_{\alpha\gamma}^{\gamma''\alpha''} |_{b'a}^{cd'}(q,r) W_{\beta\gamma}^{\gamma'\beta''} |_{cd'}^{a'b}(p,r) \end{aligned} \tag{10}$$

Quantum Inverse-Scattering Method

The Leningrad school of Faddeev incorporated the methods of Baxter and Yang in their so-called

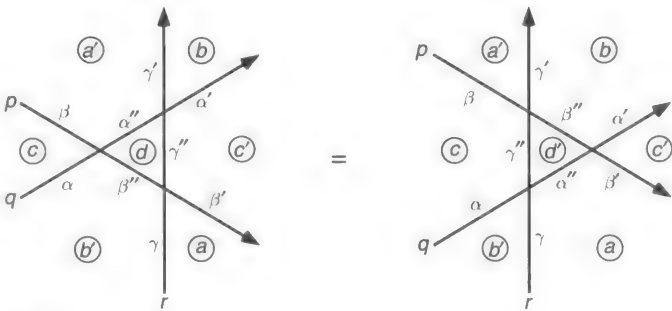


Figure 8 General YBE.

quantum inverse-scattering method (QISM), coining the term quantum YBEs (QYBEs) for eqns [8]. If special limiting values of p and q can be found, say as $\hbar \rightarrow 0$, such that $\omega_{\alpha\mu}^{\lambda\beta} = \delta_{\mu}^{\lambda} \delta_{\alpha}^{\beta} + O(\hbar)$, one can reduce [8] to the classical Yang–Baxter equations (CYBEs) by expanding up to the first nontrivial order in expansion variable \hbar . These determine the integrability of certain models of classical mechanics by the inverse-scattering method and the existence of Lax pairs.

Checkerboard generalizations

Star-triangle equations [3] and [4] imply that there are further generalizations of the YBEs, namely those for which the faces enclosed by the rapidity lines are alternately colored black and white in a checkerboard pattern. We can then introduce either vertex model weights $\omega_{\alpha\mu}^{\lambda\beta}(p,q)$ and $\bar{\omega}_{\alpha\mu}^{\lambda\beta}(p,q)$, or IRF-vertex model weights $W_{\alpha\mu|ab}^{\lambda\beta|dc}(p,q)$ and $\bar{W}_{\alpha\mu|ab}^{\lambda\beta|dc}(p,q)$, or IRF model weights $w_{ab}^{dc}(p,q)$ and $\bar{w}_{ab}^{dc}(p,q)$, see Figure 9.

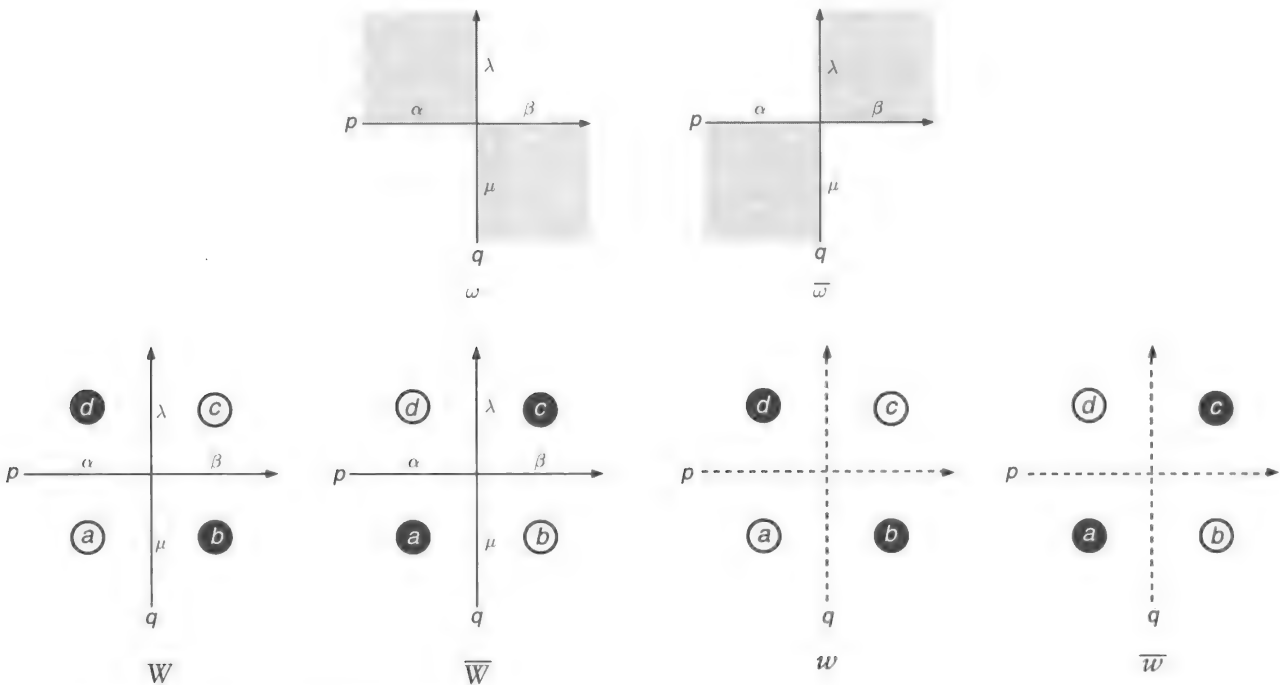


Figure 9 Checkerboard versions of the weights.

The black faces are those where the spins of the spin model with weights defined in Figure 2 live; the white faces are to be considered empty in Figures 2 and 3 (or, equivalently, they can be assumed to host trivial spins that take on only a single value). Clearly, the IRF-vertex model description contains all the other versions.

Checkerboard Vertex Model

First we consider the checkerboard vertex model with weights $\omega_{\alpha\mu}^{\lambda\beta}(p, q)$ and $\bar{\omega}_{\alpha\mu}^{\lambda\beta}(p, q)$ as assigned in Figure 9. The YBE [8] then generalizes to two sets of equations:

$$\begin{aligned} & \sum_{\alpha''} \sum_{\beta''} \sum_{\gamma''} \omega_{\beta\alpha}^{\alpha''\beta''}(p, q) \omega_{\alpha''\gamma''}^{\gamma'\alpha'}(q, r) \bar{\omega}_{\beta''\gamma}^{\gamma''\beta'}(p, r) \\ &= R(p, q, r) \sum_{\alpha''} \sum_{\beta''} \sum_{\gamma''} \bar{\omega}_{\beta''\alpha''}^{\alpha'\beta'}(p, q) \\ & \quad \times \bar{\omega}_{\alpha\gamma}^{\gamma''\alpha''}(q, r) \omega_{\beta\gamma''}^{\gamma'\beta''}(p, r) \end{aligned} \quad [11]$$

$$\begin{aligned} & \bar{R}(p, q, r) \sum_{\alpha''} \sum_{\beta''} \sum_{\gamma''} \bar{\omega}_{\beta\alpha}^{\alpha''\beta''}(p, q) \bar{\omega}_{\alpha''\gamma''}^{\gamma'\alpha'}(q, r) \omega_{\beta''\gamma}^{\gamma''\beta'}(p, r) \\ &= \sum_{\alpha''} \sum_{\beta''} \sum_{\gamma''} \omega_{\beta''\alpha''}^{\alpha'\beta'}(p, q) \omega_{\alpha\gamma}^{\gamma''\alpha''}(q, r) \bar{\omega}_{\beta\gamma''}^{\gamma'\beta''}(p, r) \end{aligned} \quad [12]$$

where scalar factors R and \bar{R} have been added as in [3] and [4]. These equations are represented graphically by Figure 10.

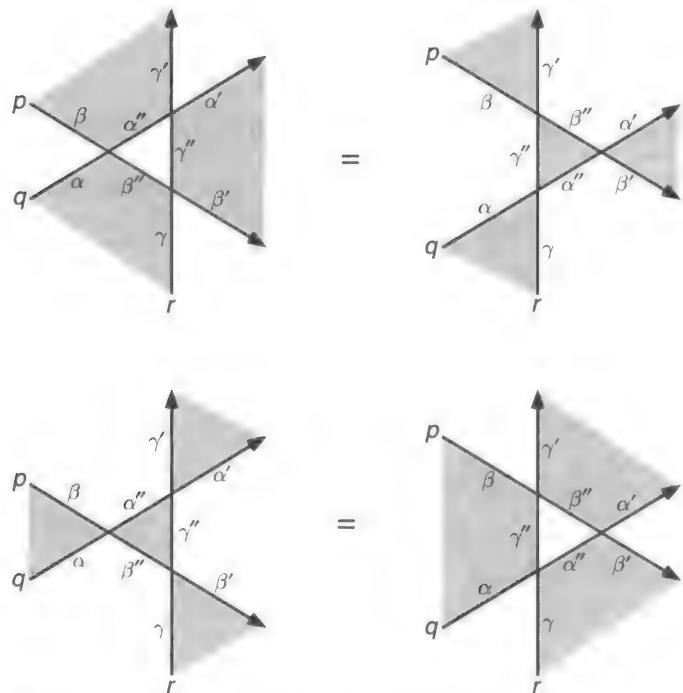


Figure 10 Checkerboard vertex model YBE.

Checkerboard IRF Model

The checkerboard IRF version of the YBE [8] becomes

$$\begin{aligned} & \sum_d w_{cb'}^{a'd}(p, q) w_{dc'}^{a'b}(q, r) \bar{w}_{b'a}^{d'c'}(p, r) \\ &= R(p, q, r) \sum_{d'} \bar{w}_{d'a}^{b'c'}(p, q) \bar{w}_{b'a}^{cd'}(q, r) w_{cd'}^{a'b}(p, r) \end{aligned} \quad [13]$$

$$\begin{aligned} & \bar{R}(p, q, r) \sum_d \bar{w}_{cb'}^{a'd}(p, q) \bar{w}_{dc'}^{a'b}(q, r) w_{b'a}^{d'c'}(p, r) \\ &= \sum_{d'} w_{d'a}^{b'c'}(p, q) w_{b'a}^{cd'}(q, r) \bar{w}_{cd'}^{a'b}(p, r) \end{aligned} \quad [14]$$

again with scalar factors R and \bar{R} added as in [3] and [4]. These equations can now be represented graphically as in Figure 11. Note that these equations reduce to eqns [3] and [4] if the spins on the white faces are allowed to take only one value, which means that they can be ignored.

Checkerboard IRF-Vertex Model

Finally, the most general case is represented by the checkerboard IRF-vertex model, with weights defined in Figure 9. For this case the YBEs are given by

$$\begin{aligned} & \sum_{\alpha''} \sum_{\beta''} \sum_{\gamma''} \sum_d W_{\beta\alpha}^{\alpha''\beta''} |^{a'd}(p, q) \\ & \quad \times W_{\alpha''\gamma''}^{\gamma'\alpha'} |^{a'b}(q, r) \bar{W}_{\beta''\gamma}^{\gamma''\beta'} |^{dc'}(p, r) \\ &= R(p, q, r) \sum_{\alpha''} \sum_{\beta''} \sum_{\gamma''} \sum_{d'} \bar{W}_{\beta''\alpha''}^{\alpha'\beta'} |^{bc'}(p, q) \\ & \quad \times \bar{W}_{\alpha\gamma}^{\gamma''\alpha''} |^{cd'}(q, r) W_{\beta\gamma''}^{\gamma'\beta''} |^{a'b}(p, r) \end{aligned} \quad [15]$$

$$\begin{aligned} & \bar{R}(p, q, r) \sum_{\alpha''} \sum_{\beta''} \sum_{\gamma''} \sum_d \bar{W}_{\beta\alpha}^{\alpha''\beta''} |^{a'd}(p, q) \\ & \quad \times \bar{W}_{\alpha''\gamma''}^{\gamma'\alpha'} |^{a'b}(q, r) W_{\beta''\gamma}^{\gamma''\beta'} |^{dc'}(p, r) \\ &= \sum_{\alpha''} \sum_{\beta''} \sum_{\gamma''} \sum_{d'} W_{\beta''\alpha''}^{\alpha'\beta'} |^{bc'}(p, q) \\ & \quad \times W_{\alpha\gamma}^{\gamma''\alpha''} |^{cd'}(q, r) \bar{W}_{\beta\gamma''}^{\gamma'\beta''} |^{a'b}(p, r) \end{aligned} \quad [16]$$

with its graphical representation in Figure 12.

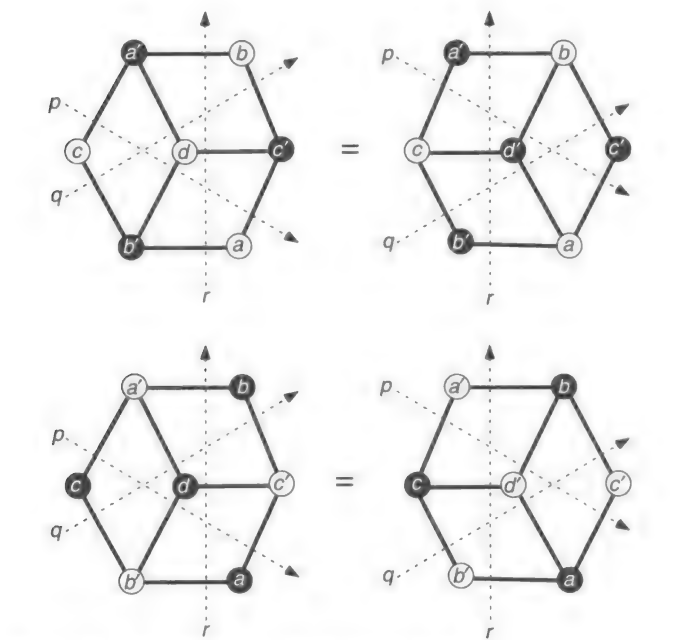


Figure 11 Checkerboard IRF model YBE.

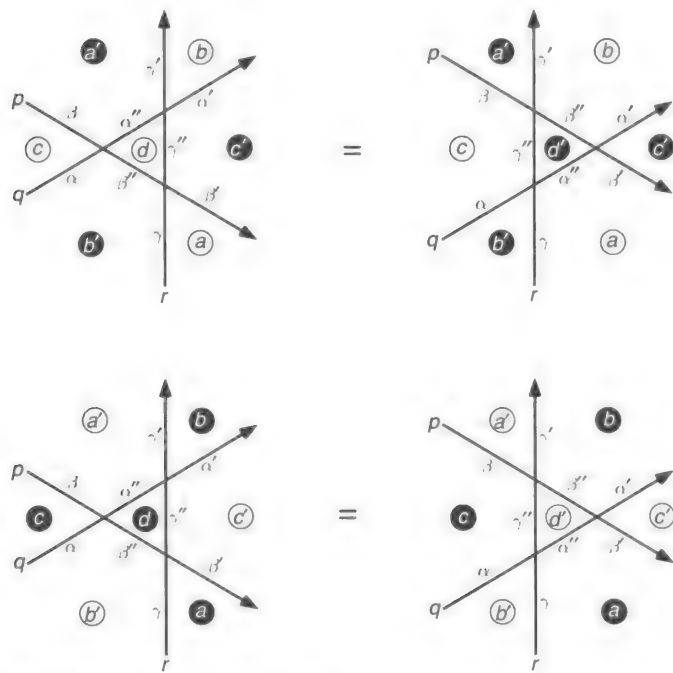


Figure 12 Checkerboard YBE.

Formal Equivalence of Languages

The Square Weight

Combining four weights of a checkerboard model in a square, as is done with four spin model weights in Figure 13, we find a regular vertex model weight with rapidities that are now pairs of the original ones. This gives

$$W_{\alpha\mu}(p_1, q_1) \overline{W}_{\mu\beta}(p_1, q_2) \overline{W}_{\alpha\lambda}(p_2, q_1) W_{\lambda\beta}(p_2, q_2) = \omega_{\alpha\mu}^{\lambda\beta}(p_1, p_2; q_1, q_2) \tag{17}$$

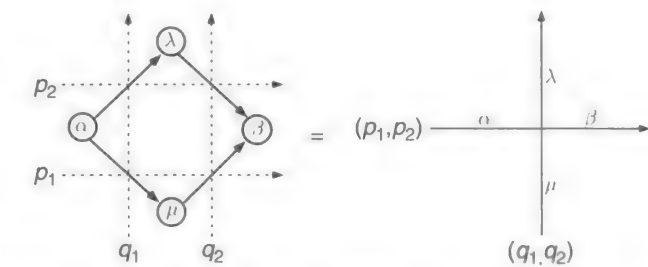


Figure 13 Square weight as vertex weight.

From any solution of [3] and [4] we can thus construct a solution of YBE [8]. This has been used by Bazhanov and Stroganov to relate the integrable chiral Potts model with a cyclic representation of the six-vertex model.

Map to Checkerboard Vertex Model

The checkerboard IRF-vertex model formulation contains all other versions mentioned above as special cases. However, collecting the state variables in triples, we can immediately translate it to a vertex model version, writing

$$\omega_{\hat{\alpha}\hat{\mu}}^{\hat{\lambda}\hat{\beta}}(p, q) = W_{\alpha\mu}^{\lambda\beta|dc}(p, q), \quad \overline{\omega}_{\hat{\alpha}\hat{\mu}}^{\hat{\lambda}\hat{\beta}}(p, q) = \overline{W}_{\alpha\mu}^{\lambda\beta|dc}(p, q) \text{ if } \begin{cases} \hat{\lambda} = (d, \lambda, c), & \hat{\beta} = (b, \beta, c) \\ \hat{\alpha} = (a, \alpha, d), & \hat{\mu} = (a, \mu, b) \end{cases} \tag{18}$$

$$\omega_{\hat{\alpha}\hat{\mu}}^{\hat{\lambda}\hat{\beta}}(p, q) = \overline{\omega}_{\hat{\alpha}\hat{\mu}}^{\hat{\lambda}\hat{\beta}}(p, q) = 0 \text{ otherwise} \tag{19}$$

In eqn [19], we have set all vertex model weights zero that are inconsistent with IRF-vertex configurations. Clearly, the translation of IRF models and spin models to vertex models can be done similarly.

Map to Spin Model

We can, furthermore, translate each vertex model with weights assigned as in Figures 6 or 9 into a spin model with weights as in Figure 2 by defining suitable spins in the black faces, after checkerboard coloring. Each spin is then defined to be the ordered set of states on the line segments of the vertex model, $\underline{a} = (\alpha_1, \alpha_2, \dots)$, ordering the line segments counterclockwise starting at, say, 12 o'clock. We can then identify $\omega_{\alpha\mu}^{\lambda\beta}(p, q) = W_{\underline{a}, \underline{b}}(p, q)$, $\overline{\omega}_{\alpha\mu}^{\lambda\beta}(p, q) = \overline{W}_{\underline{a}, \underline{b}}(p, q)$. This is surely not very economical, as many of the weights will be equal, but it helps show that all different versions of the checkerboard YBE are formally equivalent.

Hence, we shall only use the vertex model language in the following. It is fairly straightforward to convert to the other formulations.

An $sl(m|n)$ Example

One fundamental example is a Q -state model for which the rapidities have $2Q + 1$ components, $p = (p_{-Q}, \dots, p_Q)$, $q = (q_{-Q}, \dots, q_Q)$, etc., and the states on the line segments are arranged in strings of continuing conserved color. The vertex weights, for $\alpha, \beta, \lambda, \mu = 1, \dots, Q$, are given by

$$\omega_{\alpha\mu}^{\lambda\beta}(p, q) = \omega_{0\alpha\mu}^{\lambda\beta}(p_0, q_0) \frac{p_{+\lambda} q_{-\beta}}{q_{+\alpha} p_{-\mu}} \tag{20}$$

with $(\rho \neq \sigma)$

$$\begin{aligned} \omega_{0\rho\rho}^{\rho\rho}(p_0, q_0) &= \mathcal{N} \sinh[\eta + \varepsilon_\rho(p_0 - q_0)] \\ \omega_{0\sigma\rho}^{\rho\sigma}(p_0, q_0) &= \mathcal{N} G_{\rho\sigma} \sinh(p_0 - q_0) \\ \omega_{0\sigma\rho}^{\sigma\rho}(p_0, q_0) &= \mathcal{N} e^{(p_0 - q_0)\text{sign}(\rho - \sigma)} \sinh \eta \\ \omega_{0\alpha\mu}^{\lambda\beta}(p_0, q_0) &= 0, \quad \text{otherwise} \end{aligned} \tag{21}$$

where \mathcal{N} is an arbitrary overall normalization factor and η is a constant. Furthermore, $\varepsilon_\rho = \pm 1$ for $\rho = 1, \dots, Q$, where m of them equal $+1$ and n of them equal -1 . The $G_{\rho\sigma}$'s are constants satisfying $G_{\rho\sigma} = 1/G_{\sigma\rho}$, which freedom is allowed because the number of ρ - σ crossings minus the number of σ - ρ crossings is fixed by the states on the boundary only, that is, the choice of $\alpha, \alpha', \beta, \beta', \gamma, \gamma'$ in YBE [8] and Figure 5.

The solution [20], [21] has many applications. The case $m = 0, n = 2$ leads to the general six-vertex model; the $m = 0, n = n$ case produces the fundamental intertwiner of affine quantum group $U_q \widehat{sl}(n)$, whereas the case $m = 2, n = 1$ corresponds to the supersymmetric one-dimensional t - J model.

Operator Formulations

The R -Matrix

For a problem with N rapidity lines, carrying rapidities p_1, \dots, p_N , we can introduce a set of matrices $R_{ij}(p_i, p_j)$, for $1 \leq i < j \leq N$, with elements

$$R_{ij}(p_i, p_j)_{\alpha_1 \dots \alpha_N}^{\beta_1 \dots \beta_N} = \omega_{\alpha_i \alpha_j}^{\beta_i \beta_j}(p_i, p_j) \prod_{k \neq i, j} \delta_{\alpha_k}^{\beta_k} \tag{22}$$

In terms of these, the YBE [8] can be rewritten in matrix form as

$$\begin{aligned} R_{jk}(p_j, p_k) R_{ik}(p_i, p_k) R_{ij}(p_i, p_j) \\ = R_{ij}(p_i, p_j) R_{ik}(p_i, p_k) R_{jk}(p_j, p_k) \end{aligned} \tag{23}$$

where $1 \leq i < j < k \leq N$.

The \check{R} -Matrix

If we transpose the β indices β_i and β_j in eqn [22], we can define a set of matrices $\check{R}_{i, i+1}(p, q)$ with elements

$$\check{R}_{i, i+1}(p, q)_{\alpha_1 \dots \alpha_N}^{\beta_1 \dots \beta_N} = \omega_{\alpha_i, \alpha_{i+1}}^{\beta_i, \beta_{i+1}}(p, q) \prod_{k \neq i, i+1} \delta_{\alpha_k}^{\beta_k} \tag{24}$$

Using these, the YBE [8] can be rewritten in matrix form as

$$\begin{aligned} \check{R}_{i, i+1}(q, r) \check{R}_{i+1, i+2}(p, r) \check{R}_{i, i+1}(p, q) \\ = \check{R}_{i+1, i+2}(p, q) \check{R}_{i, i+1}(p, r) \check{R}_{i+1, i+2}(q, r) \end{aligned} \tag{25}$$

and

$$[\check{R}_{i, i+1}(p, q), \check{R}_{j, j+1}(r, s)] = 0, \quad \text{if } |i - j| \geq 2 \tag{26}$$

In this formulation, it is clear that many solutions can be found “Baxterizing” Temperley–Lieb and Iwahori–Hecke algebras.

Classical YBEs

If we expand

$$R_{ij}(p_i, p_j) = 1 + \hbar X_{ij}(p_i, p_j) + O(\hbar^2) \tag{27}$$

in [23], we get in second order in \hbar the classical YBE (CYBE) as the vanishing of a sum of three commutators, that is,

$$\begin{aligned} [X_{ij}(p_i, p_j), X_{ik}(p_i, p_k)] + [X_{ij}(p_i, p_j), X_{jk}(p_j, p_k)] \\ + [X_{ik}(p_i, p_k), X_{jk}(p_j, p_k)] = 0 \end{aligned} \tag{28}$$

introduced by Belavin and Drinfel'd, where X_{ij} is called the classical r -matrix.

Reflection YBEs

Cherednik and Sklyanin found a condition determining the solvability of systems with boundaries, the reflection YBEs (RYBEs), see Figure 14. Upon

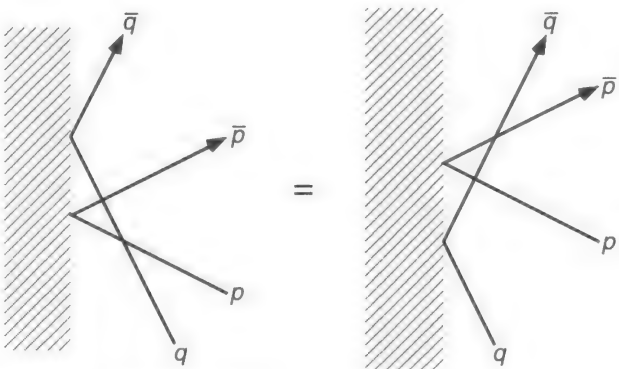


Figure 14 Reflection YBE.

collisions with a left or right wall the rapidity variable changes from p to \bar{p} and back. In most examples, in which the rapidities are difference variables such that $R(p, q) = R(p - q)$, one also has $\bar{p} = \mu - p$, with μ some constant. The corresponding left boundary weights are $K_\alpha^\beta(p, \bar{p})$ satisfying

$$\begin{aligned} \check{K}_1(q, \bar{q}) \check{R}_{12}(\bar{p}, q) \check{K}_1(p, \bar{p}) \check{R}_{12}(q, p) \\ = \check{R}_{12}(\bar{p}, \bar{q}) \check{K}_1(p, \bar{p}) \check{R}_{12}(\bar{q}, p) \check{K}_1(q, \bar{q}) \end{aligned} \quad [29]$$

with $\check{K}_1(p, \bar{p})$ defined by a direct product as in [24] appending unit matrices for positions $i \geq 2$, and a similar equation must hold for the right boundary. Most work has been done for vertex models, while Pearce and co-workers wrote several papers on the IRF-model version.

Higher-Dimensional Generalizations

In 1980 Zamolodchikov introduced a three-dimensional generalization of the YBE, the so-called tetrahedron equations (TEs), and he found a special solution. Baxter then succeeded in proving that this solution satisfies all TEs. Baxter and Bazhanov showed in 1992 that this solution can be seen as a special case of the $sl(\infty)$ chiral Potts model. Several authors found further generalizations more recently.

Inversion Relations

When $\omega_{\alpha\mu}^{\lambda\beta}(p, p) \propto \delta_\alpha^\lambda \delta_\mu^\beta$, that is, the weight decouples when the two rapidities are equal, one can derive the local inverse relation depicted in Figure 15, which is a generalization of the Reidemeister move of type II in Figure 4. It is easily shown that $C(q, p) = C(p, q)$.

This local relation implies also a global inversion relation which can be found in many ways. The following heuristic way is the easiest: consider the situation in Figure 16, with N closed p -rapidity lines and M closed q -rapidity lines. For M and N large, we may expect the partition function of Figure 16 to factor asymptotically in top- and bottom-half contributions. If each line segment carries a state

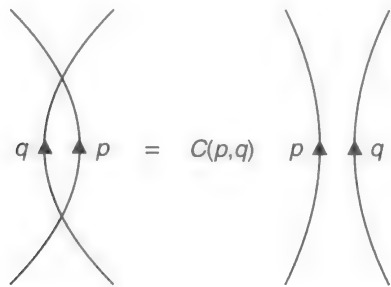


Figure 15 Local inversion relation.

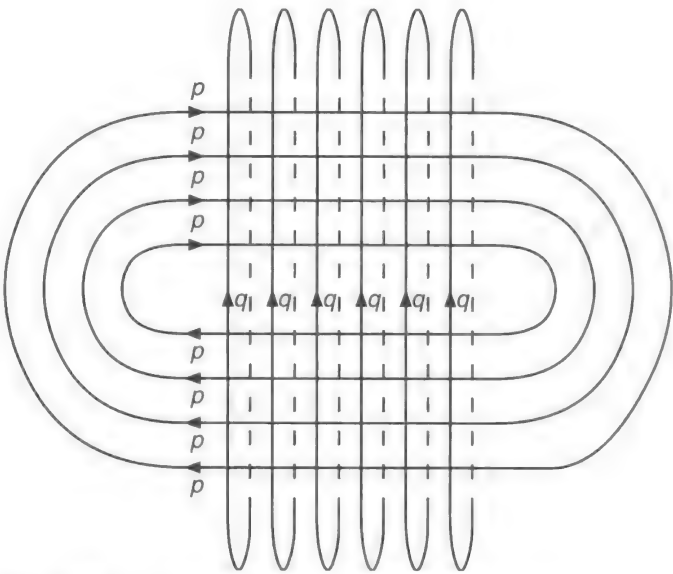


Figure 16 Heuristic derivation of inversion relation.

variable that can assume Q values, then the total partition function factors by repeated application of the relation in Figure 15 into the contribution of $M + N$ circles. Therefore,

$$Z = Q^{M+N} C(p, q)^{MN} \approx Z_{M,N}(p, q) Z_{N,M}(q, p) \quad [30]$$

Taking the thermodynamic limit,

$$z(p, q) \equiv \lim_{M, N \rightarrow \infty} Z_{M,N}(p, q)^{1/MN} \quad [31]$$

one finds

$$z(p, q) z(q, p) = C(q, p) \quad [32]$$

In many models, eqn [32], supplemented with some suitable symmetry and analyticity conditions, can be used to calculate the free energy per site.

See also: Affine Quantum Groups; Bethe Ansatz; Classical r -matrices, Lie Bialgebras, and Poisson Lie Groups; Eight Vertex and Hard Hexagon Models; Hopf Algebras and q -Deformation Quantum Groups; Integrability and Quantum Field Theory; Integrable Discrete Systems; Integrable Systems: Overview; The Jones Polynomial; Knot Invariants and Quantum Gravity; Knot Theory and Physics; Sine-Gordon Equation; Topological Knot Theory and Macroscopic Physics; Two-Dimensional Ising Model; von Neumann Algebras; Subfactor Theory.

Further Reading

Au-Yang H and Perk JHH (1989) Onsager’s star-triangle equation: master key to integrability. *Advanced Studies in Pure Mathematics* 19: 57–94.
Baxter RJ (1982) *Exactly Solved Models in Statistical Mechanics*. London: Academic Press.

- Behrend RE, Pearce PA, and O'Brien DL (1996) Interaction-round-a-face models with fixed boundary conditions: the ABF fusion hierarchy. *Journal of Statistical Physics* 84: 1–48.
- Gaudin M (1983) *La Fonction d'Onde de Bethe*. Paris: Masson.
- Jimbo M (ed.) (1987) *Yang-Baxter Equation in Integrable Systems*. Singapore: World Scientific.
- Kennelly AE (1899) The equivalence of triangles and three-pointed stars in conducting networks. *Electrical World and Engineer* 34: 413–414.
- Korepin VE, Bogoliubov NM, and Izergin AG (1993) *Quantum Inverse Scattering Method and Correlation Functions*. Cambridge: Cambridge University Press.
- Kulish PP and Sklyanin EK (1981) Quantum spectral transform method. Recent developments. In: Hietarinta J and Montonen C (eds.) *Integrable Quantum Field Theories*, Lecture Notes in Physics, vol. 151, pp. 61–119. Berlin: Springer.
- Lieb EH and Wu FY (1972) Two-dimensional ferroelectric models. In: Domb C and Green MS (eds.) *Phase Transitions and Critical Phenomena*, vol. 1, pp. 331–490. London: Academic Press.
- Onsager L (1971) The Ising model in two dimensions. In: Mills RE, Ascher E, and Jaffee RI (eds.) *Critical Phenomena in Alloys, Magnets and Superconductors*, pp. xix–xxiv, 3–12. New York: McGraw-Hill.
- Perk JHH (1989) Star-triangle equations, quantum Lax pairs, and higher genus curves. *Proceedings of Symposia in Pure Mathematics* 49(1): 341–354.
- Perk JHH and Schultz CL (1981) New families of commuting transfer matrices in q -state vertex models. *Physics Letters A* 84: 407–410.
- Perk JHH and Wu FY (1986) Graphical approach to the nonintersecting string model: star-triangle equation, inversion relation, and exact solution. *Physica A* 138: 100–124.
- Reidemeister K (1926a) Knoten und Gruppen. *Abhandlungen aus dem Mathematischen Seminar der Hamburgischen Universität* 5: 7–23.
- Reidemeister K (1926b) Elementare Begründung der Knotentheorie. *Abhandlungen aus dem Mathematischen Seminar der Hamburgischen Universität* 5: 24–32.
- Yang CN (1967) Some exact results for the many-body problem in one dimension with repulsive delta-function interaction. *Physical Review Letters* 19: 1312–1314.
- Yang CN (1968) S-matrix for the one-dimensional N-body problem with repulsive or attractive δ -function interaction. *Physical Review* 167: 1920–1923.

INDEX

This index is in letter-by-letter order, whereby hyphens and spaces within index headings are ignored in the alphabetization, and it is arranged in set-out style, with a maximum of three levels of heading. Location references refer to the volume number (in bold) and page number (separated by a colon). Major discussion of a subject is indicated by a bold page range.

A

- A–B effect *see* Aharonov–Bohm effect
- Abel equations, and singularity theory 4:592
- Abelian BF theories
 - and linking numbers 1:262
 - quantization 1:260
 - and Ray–Singer torsion 1:260
 - special features 1:259
- Abelian chiral anomalies, characteristics 1:205
- Abelian duality, and Wu–Yang criterion 2:204
- Abelian gauge theories
 - characteristics 3:575
 - with differential forms 1:141–150
 - as infinite-dimensional Lie group example 3:40
- Abelian Higgs model
 - critical coupling 1:156
 - interaction energy 1:153
 - overview 1:151
 - radially symmetric vortices 1:152
 - research developments 1:159
 - useful abbreviations 1:159
 - vortex characteristics 4:605
 - vortex overview 1:151–160
 - vortex scattering 1:155
- Abnormal fluctuations, and quantum
 - central-limit theorems 4:136
- Absolute continuity, and measures on loop space 3:416
- Absolute Galois group, and conformal field theory 3:604
- AC *see* Asymptotic causality
- ACI *see* Algebraically completely integrable
- Action–angle coordinates, classical
 - mechanics overview 1:7
- Action-angle transforms, in
 - Calogero–Moser–Sutherland systems 1:409
- Action functional, and Hamilton–Jacobi equations 2:642
- Adaptive discretization, and wavelet computation 5:417
- Adaptive strategy, and wavelet computation 5:418
- Adaptive wavelet scheme, and turbulence 5:417
- Additivity
 - and entanglement measures 2:235
 - in *K*-theory 3:247
 - in twisted *K*-theory 3:253
- Adele classes
 - characteristics 3:607
 - and class field theory 3:603
 - and motives 3:607
- ADHM formalism *see*
 - Atiyah–Drinfeld–Hitchin–Manin formalism
- Adiabatic Euler equations, inviscid
 - compressible fluid flows 1:599
- Adiabatic evolution
 - overview 2:530
 - piston 1:165
- Adiabatic pistons
 - exact equations 1:163
 - with finite mass 1:167
 - infinite cylinder distribution 1:164
 - as macroscopic problem 1:160
 - microscopical model 1:162
 - as microscopic problem 1:160
 - and multiscale approaches 3:476
 - numerical simulations 1:169
 - open problems 1:171
 - overview 1:160–174
 - thermodynamic limit 1:165
- Adjoint operator, spectral theory for 4:634
- Adler–Kostant–Symes method
 - for finite open-end Toda lattice 5:239
 - and symplectic reduction 3:74
- ADM mass *see* Arnowitt–Deser–Misner mass
- AdS *see* Anti-de Sitter spacetime
- AdS–CFT correspondence *see* Anti-de Sitter–conformal field theory correspondence
- Advanced kernels, and microcausality 2:95
- AF *see* Antiferromagnetism
- AF algebras *see* Approximately finite-dimensional C^* -algebras
- Affine Kac–Moody algebras, *q*-special
 - function applications 4:116
- Affine Lie algebras
 - and solitons 4:596, 4:599
 - for Toda lattices 5:241
 - as transformation groups 4:598
- Affine quantum groups
 - boundary quantum groups 1:189
 - integrable quantum spin chains 1:189
 - overview 1:183–190
- symmetry 1:188
- Yangians 1:186
- Affine Toda theories
 - and classical integrable systems 3:51
 - and integrable QFT 3:55
- Aggregation techniques, and multiscale
 - approaches 3:473
- Aharonov–Anandan phase,
 - overview 2:532
- Aharonov–Bohm effect
 - and A–B geometry 1:195
 - basic experiment 1:194
 - covariant derivative 1:196
 - flat connections 1:196
 - holonomy 1:196
 - nonabelian and gravitational effects 1:198
 - overview 1:191–198, 1:195
 - parallel transport 1:196
 - Schrödinger equation solution 1:193
- AHM *see* Abelian Higgs model
- Aircraft trailing vortices, properties 5:393
- Airy process, in random matrix
 - theory 2:590
- AKS method *see* Adler–Kostant–Symes method
- AKSZ formalism, and graded Poisson algebras 2:566
- Algebra of constraints, and Einstein gravity 5:456
- Algebra of functions, noncommutative
 - tori 3:524
- Algebraically completely integrable
 - and duality, Fourier–Mukai transform, bispectrality 3:77
 - and special functions 3:76
- Algebraic geometry
 - and complex analysis 4:548
 - and integrable systems, overview 3:65–78
 - and multidimensional integrable lattices 3:81
- Algebraic intersection products
 - normal cone deformation 3:152
 - refined intersections 3:152
- Algebraic quantum field theory
 - fields as natural transformations 1:203
 - general concepts 5:329
 - idealized localizations 1:201
 - and Jones polynomial 3:186
 - overview 1:198–204
 - particles 1:202

- Algebraic quantum field theory (*continued*)
 phase space analysis 1:201
 sectors 1:202
 statistics 1:202
 and subfactors 5:388
 symmetries 1:202
 and Tomita–Takesaki modular theory 5:256
 traditional approach 1:200
 ultraviolet structure 1:201
- Algebraic topology
 characteristics 1:139
 graded Poisson algebra examples 2:564
- Algebras
see also Lie algebras
 associative algebras 2:314
 CCR algebra 4:132
 and channels 1:472
 chord diagrams 3:232
 current algebras 1:674–679
 q -deformation coordinate algebras 2:692
 q -deformation enveloping algebras 2:690
 differential algebras 2:243
 differential structures 4:237
 extended algebras 1:339
 finite-dimensional algebras 2:313–322
 first-order deformation 2:18
 formal deformation 2:19
 homological algebra 1:386
 Hopf algebras 1:355, 2:687
 integrals of motion, and recursion operators 3:91
 Kontsevich integral 3:233
 operads in 3:611
 operator algebras 1:95
 Poincaré algebra 2:428
 quantized universal enveloping, and deformations 2:19
 quantum affine algebras 1:184
 and quantum dynamical semigroups 4:164
 quasi-Hopf algebras 2:697
 spinor algebra 4:668
 Steenrod algebra 1:548
 Sugawara algebra 1:676
 for supermanifolds 5:129
 Temperley–Lieb algebra 3:181
 in topologies 1:139
 Verlinde algebra 5:114
 vertex operator, and conformal field theory 5:319
 Virasoro algebra 1:292–294, 1:334, 1:676
 von Neumann algebras 5:379–384
 Weil algebra 2:244
- C^∞ -Algebras
 and Atiyah–Singer index theorem 3:250
 and Baum–Connes conjecture 3:250
 commutative 4:89
 GNS construction 4:91
 K -theory 3:249
 normal and singular states 4:91
 overview 1:393–398
 positive cones in 4:88
 positive maps on, overview 4:88–94
 vector states and density matrices 4:90
- *-Algebras, construction 4:203
- A_∞ -Algebras, operads in 3:611
- L_∞ -Algebras, operads in 3:611
- Algorithms
 Dirac, and constrained systems 1:612
 message-passing algorithms 5:53
 for multicomponent fluid flow 3:140
 multigrid algorithms 3:474
- Almost-Mathieu operator, and quasiperiodic potentials 3:335
- Alternating-sign matrices, in combinatorics 1:572
- Aluminum, industrial production 3:381
- Amalgamation, over subalgebra, free independence 2:418
- Amplitude equations
 and multiscale approaches 3:471
 wave propagation inverse problems 1:344
- Analytic classification theory, subfactors 5:387
- Analytic continuation, first theorems 4:541
- Analyticity properties
 overview 4:484
 and scattering in relativistic QFT 4:465–475
- Analytic N -point functions, in massive field theory 4:478
- Anderson localisation, and quantum Hall effect 4:248
- Angular momentum, and symplectic reduction 5:192
- Animal electricity *see* Galvanism
- Annulus partition function, in boundary CFT 1:336
- Anomalies
 and Atiyah–Singer index theorem 1:211
 chiral anomalies 1:205
 covariant *vs.* consistent 1:210
 in lattice gauge theories 1:211
 overview 1:205–212
 renormalization 4:405
 in spacetime dimensions 1:210
 in string theory 1:210
 trace anomalies 1:209
- Anomalous diffusion, in fractal medium 3:480
- Anosov diffeomorphisms, as hyperbolic dynamical system 2:724
- Antiautomorphisms, Clifford algebras 1:521
- Anticommuting variables, in supersymmetry 5:146
- Anti-de Sitter–conformal field theory
 correspondence
 bulk–boundary correspondence 1:178
 coincident D3 branes 1:175
 confining theory path 1:182
 correlation functions 1:178
 Dirichlet brane geometry 1:175
 and Einstein manifolds 1:180
 essential characteristics 1:177
 geometrical considerations 1:174
 and holography 5:139
 operator dimensions 1:178
 overview 1:174–183
 and thermal gauge theory 1:176
 two-point functions 1:178
 Wilson loop calculations 1:179
- Anti-de Sitter spacetime
 definition 1:382
 and Einstein equations solutions 2:166
- Antiferromagnetism, and superconductivity unification 2:648
- Antifields, and Batalin–Vilkovisky quantization 1:249
- Anti-self-dual Yang–Mills equations
 dimensional reductions 3:47
 in twistor theory 5:306
- Antisolitons, and sine-Gordon equation 4:578
- Applied magnetic field, and Ginzburg–Landau equations
 Meissner solutions 2:549
 overview 2:549
 superconductivity nucleation 2:550
 vortices solutions 2:550
- Approximately finite-dimensional C^* -algebras, characteristics 1:395
- Approximations
 Boussinesq approximations 3:493
 in fluid mechanics 2:369
 Galerkin approximation 5:376
 mean field approximations 5:452
 minisuperspace approximation 4:154
 quasistationary approximation 3:472
 semiclassical approximation 3:578–585
 stationary phase approximation 5:44–49
 and wavelets 5:433
 WKB approximations 2:442
 zero-mode approximation 5:159
- AQC *see* Arithmetic quantum chaos
- AQFT *see* Algebraic quantum field theory
- Arakelov geometry, characteristics 3:605
- Arbitrary spin, relativistic wave equations 4:396
- Area
 and holomorphic dynamics 2:658
 and open string field theory 5:99
- Area formula, and geometric measure theory 2:523

- Area-minimizing surfaces,
construction 3:428
- Arhus integral, Le–Murakami–Ohtsuki
invariant 2:352
- Arithmetic algebraic geometry,
types 3:604
- Arithmetical physics, speculations 3:607
- Arithmetic Calabi–Yau geometry,
characteristics 3:604
- Arithmetic intersection theory,
overview 3:155
- Arithmetic quantum chaos
cat map quantum eigenstates 1:219
closed geodesics 1:216
eigenfunctions 1:218
eigenvalue statistics 1:214
Hecke operators 1:217
and hyperbolic surfaces 1:213
 L -functions 1:218
overview 1:212–221
and quantum unique ergodicity 1:216
Selberg trace formula 1:214
- Arnol'd diffusion
classical mechanics overview 1:26
and Nekhoroshev's theorem 5:29
stability theory 4:592
- Arnol'd–Liouville theorem, classical
mechanics overview 1:8
- Arnol'd nondegeneracy, and Hamiltonian
system stability 2:634
- Arnol'd theorem
and Floer homology 2:358
general case 3:197
and Hamiltonian system stability 2:634
overview 3:194
reformulation for study 2:359
- Arnowitt–Deser–Misner mass
definition 1:382
and general relativity 1:414
- Artificial magnetism, and negative
refraction 3:485
- Artificial plasma, and negative
refraction 3:484
- Ashtekar formalism
general relativity 1:416
and knot invariants 3:217
- Askey–Wilson polynomials
continuous q -ultraspherical polynomial
as 4:110
overview 4:110
- Associative algebras, overview 2:314
- Asymmetric cloning, quantum states 3:631
- Asymptotically flat spacetime,
definition 1:223
- Asymptotically simple spacetime,
definition 1:223
- Asymptotical models, in
hydrodynamics 2:133
- Asymptotic behavior, generalized Burgers
equations solutions 1:450
- Asymptotic causality
in low-energy region 4:485
in massive field theory 4:479
overview 4:484
and particles and Landau
singularities 4:479
and structure equations 4:486
- Asymptotic completeness
and QFT analytic program 4:473
and scattering theory 4:464
- Asymptotic conversion, and comparative
entanglement 2:234
- Asymptotic counting formulas, in random
algebraic geometry 4:327
- Asymptotic dynamics, thermal fields 5:234
- Asymptotic eigenfunctions, and quantum
ergodicity 4:186
- Asymptotic expansion, loop integrals 4:37
- Asymptotic freedom
in constructive QFT 1:623
and perturbative quantum
chromodynamics 4:149
quantum chromodynamics 4:147
- Asymptotic freeness, random
matrices 2:421
- Asymptotic information, from generating
functions 1:560
- Asymptotic lattices, and reductions 3:85
- Asymptotic Lyapunov stability, fluid
flow 5:4
- Asymptotic particle counters, in scattering
theory 4:460
- Asymptotic stability, and equivariant
bifurcation theory 5:188
- Asymptotic structure
asymptotically flat spacetimes 1:223
BMS group 1:224
Bondi–Sachs mass-loss formula 1:225
and caustics 3:624
Minkowski spacetime infinity 1:222
overview 1:221–226
Peeling property 1:224
- Asymptotic equipartition, and classical
data compression 4:610
- Asymptotic velocity, in random walks in
random environments 4:356
- Atiyah–Drinfeld–Hitchin–Manin
formalism
construction in gauge theory 2:472
and instantons 3:46
and noncommutative
instantons 3:522–523
- Atiyah–Hirzebruch spectral sequence,
properties 4:631
- Atiyah–Jeffrey formula, and
Mathai–Quillen formalism 3:396
- Atiyah–Singer index theorem
anomaly relationship 1:211
 C^* -algebra generalization 3:250
overview 3:248
- Atmospheric primitive equations
boundary conditions 2:538
overview 2:537
pseudogeometrical domain 2:538
vertical coordinate change 2:538
- Atmospheric vortex, characteristics 5:395
- Atoms
dimensions 1:62
equilibrium statistical mechanics
overview 1:51
- Attraction, coalescence catastrophe from
1:84
- Attractors
and chaos 1:477–488
overview 4:323–329
in random algebraic geometry 4:325
- Aubry sets, and Hamilton–Jacobi
equations 2:640
- Austenite–Martensite transformations, and
microstructures 5:363
- Automorphisms
and bounded domains 4:542
in Fedosov quantization 2:295
quantum group bundles 4:242
- Averaging lemma, and kinetic
equations 3:206
- Averaging methods
applications 1:231
averaging principle 1:226
Bogolyubov's system 1:230
fast variable elimination 1:228
justification 1:229
and multiscale approaches 3:472
overview 1:226, 1:226–231
in slow–fast systems 1:230
- Averaging stability, and Hamiltonian
systems 2:631
- Axial isotropy subgroups, and equivariant
bifurcation theory 5:187
- Axiomatic quantum field theory
interacting fields 1:239
overview 1:234–240, 3:20
PCT symmetry 1:238
 P_1 CT symmetry 1:238
quantum energy inequalities 1:239
quantum fields in 1:235
Reeh–Schlieder theorem 1:238
spin and statistics 1:238
Unruh effect 1:238
vacuum dynamical property 1:238
- ## B
- Bäcklund transformations
in discretization 1:246
and integrable PDEs 3:117
overview 1:241–247
and Painlevé equations 4:3
and solitons 4:596
and symmetries 1:245

- Bäcklund transformations (*continued*)
 via integrability 1:242
- Bailey's transformation, terminating
 balanced $_{10}W_9$ 4:109
- Bailey's $_{10}W_9$ transformation formula,
 elliptic analog 4:115
- Baker–Akiezer functions, for integrable
 systems 2:431
- Balanced categories *see* Ribbon categories
- Balanced random walks,
 characteristics 4:369
- Ball–box systems, and solitons 4:601
- Balmer–Rydberg formula, and standard
 model 3:509
- Banach space
 and spectra theory 4:636
 symmetric functionals 3:331
- Band matrices, and Toda lattices 5:243
- Bardeen–Cooper–Schrieffer–Leggett pair
 condensation, and superfluids 5:118
- Bardeen–Cooper–Schrieffer theory,
 overview 2:645
- Barycentric coordinates, in fluid
 mechanics 2:368
- Baryon–antibaryon asymmetry,
 properties 5:170
- Baryon numbers, and symmetries 5:169
- Base point free theorem, and compact
 complex manifolds 4:553
- Basis sets, orbits 1:483
- Batalin–Fradkin–Vilkovisky construction,
 and operads 3:614
- Batalin–Vilkovisky algebras
 construction, and operads 3:614
 definition 2:561
- Batalin–Vilkovisky quantization
 antifields 1:249
 classical master equation 1:249
 and gauge fixing 1:250
 gauge transformation set 1:247
 overview 1:247–253
 and Poisson–sigma model 1:252
 and topological field theories 1:251
- Baum–Connes conjecture, C^* -algebra
 generalization 3:250
- Bayesian framework, for image
 processing 3:2
- B-boundary, and spacetime
 boundaries 1:328
- BCFT *see* Boundary conformal field theory
- BC method *see* Boundary control method
- BCS *see* Bardeen–Cooper–Schrieffer theory
- BEC *see* Bose–Einstein condensation
- Bell's theorem, and quantum mechanics
 foundation 4:262
- Beltrami equation, incompressible perfect
 fluid 3:161
- Bénard's problem, in fluid dynamics 1:281
- Benjamin–Bona–Mahoney equation,
 overview 3:245
- b -equation
 dispersionless, pulson solutions 4:16
 and shallow-water background for
 peakons 4:13
- Berezin integral, in Mathai–Quillen
 construction 3:390
- Berezin-type star products, on Hermitian
 symmetric spaces 2:31
- Berry's phase, overview 2:531
- Bessel function
 characteristics 3:638
 example 3:639
 modified function 3:639
- q -Bessel function, equations 4:111
- Bessel potentials, Sobolev spaces 2:218
- Beta function, characteristics 3:637
- q -Beta functions, in q -hypergeometric
 series 4:107
- Bethe ansatz
 and factorizable S -matrices 5:466
 overview 1:253–257
 sector diagonalization 1:254
- B -field, open-string system 3:515
- BF theories
 basic classical properties 1:258
 constrained quantum, spin foam gravity
 models 4:652
 and linking numbers 1:262
 overview 1:257–264
 quantization 1:260
 and Ray–Singer torsion 1:260
 special features 1:259
- Bialgebras
 Hochschild cohomology 2:685
 Lie bialgebras 1:511–518
- Bianchi identity, and Einstein–Cartan
 theory of gravitation 2:193
- Bianchi models, reduction 2:618
- Bicrossproduct Hopf algebras
 and cogravity 1:265
 and matched pair equations 1:269
 overview 1:265–274
- Bicrossproduct Poincaré quantum groups,
 and noncommutative spacetime 1:272
- Bidimensional turbulence, associated
 statistics 5:77
- Bidirectional identical synchronization,
 characteristics 5:218
- Bifurcation theory
 in fluid dynamics, overview 1:281–285
 global 1:280
 and Leray–Schauder theory 3:284
 local 1:277
 one and two dimension examples 1:275
 overview 1:275–281
 and periodic orbits, overview 1:285–290
 from relative equilibria and periodic
 solutions 5:189
 and semi-classical trace formula 4:517
 and singularities
- fast–slow systems 4:593
- homoclinic bifurcation 4:592
- normal forms 4:591
- overview 4:588–594
- periodic orbit bifurcations 4:592
- pitchfork bifurcation 4:590
- Poincaré center-focus problem 4:592
- and stability theories 4:592
- transcritical bifurcation 4:590
- Big Bang, quantum geometry
 applications 4:235
- Big q -Jacobi polynomials,
 characteristics 4:111
- Bi-Hamiltonian systems
 in infinite Toda lattice 5:237
 iterative properties and
 integrability 3:463
 and Korteweg–de Vries equation 1:292
 overview 3:459
 soliton theory examples 1:294
 soliton theory overview 1:290–296
- Biharmonic problem, and linear
 theory 2:222
- Bilayers, and fractional quantum Hall
 effect 2:410
- Bilinear sum, Macdonald
 polynomials 4:114
- Billiard ball map, in billiard system 1:296
- Billiard flow, in billiard system 1:296
- Billiard reflection law, in billiard
 system 1:296
- Billiard system, in bounded convex
 domains
 caustics 1:297
 integrable billiards 1:298
 length spectrum 1:299
 overview 1:296–300
 periodic orbits 1:298
- Billiard trajectories
 characteristics 4:84
 complexity 4:85
 periodic types 4:85
- BIMs *see* Boundary-integral methods
- Binary pulsars, general relativity
 tests 2:486
- Binary systems, and gravitational
 waves 2:584
- Bingham model, and yield stress
 fluids 3:565
- q -Binomial series, equations 4:106
- Binomial series, example 3:639
- Binomial sums, in combinatorics 1:573
- Bipartite graphs, in combinatorics 2:62
- Birkhoff normal forms
 applicable calculus 3:581
 Birkhoff's theorem 3:579
 and mode interactions 5:189
 proof scheme 3:580
 and quantum completely integrable
 system eigenfunctions 2:150

- Birkhoff series, classical mechanics
overview 1:18
- Birkhoff's theorem, and normal
form 3:579
- Bismut–Chern character, and path
integrals 4:10
- Bispectrality, and algebraically completely
integrable 3:77
- Black branes, and black hole statistical
mechanics
entropy of extremal three-charge system
1:378
microscopic entropy for two-charge
systems 1:377
overview 1:377
- Black holes
black p -brane solutions 1:374
brane overview 1:373–382
brane solutions in string theory 1:374
BTZ black hole 4:501
and D- and black branes 1:377
 $D1$ – $D5$ – N system 1:375
duality and branes 1:375
dynamical situations 1:304
and Einstein's equations 2:68
FRW–TOV shock matching 4:565
in general relativity 2:512
global equilibrium 1:300
and gravitational waves 2:584–585
higher-derivative term brane
effects 1:380
local equilibrium 1:302
microscopic models 5:139
microstate brane geometry 1:381
and minimal surfaces 3:430
outstanding brane problems 1:381
overview 1:300–305
quantum geometry applications 4:233
Reissner–Nordstrom black hole
1:382
and spherically symmetric EDYM
equations 2:70
stationary 5:38–44
- Bloch equation
and magnetic resonance imaging 3:367
and quantum dynamical
semigroups 4:160
- Block matrix decomposition, Lie
groups 3:299
- Block spin renormalization group,
setup 4:408
- Blow-up point, quantum completely
integrable systems 2:149
- Blow-up problem, and incompressible
Euler equations 3:12
- BMS group *see* Bondi–Metzner–Sachs
group
- B nonconservation, and
Glashow–Weinberg–Salam
theory 2:213
- Bogolyubov's systems
averaging in 1:230
inequality in equilibrium statistical
mechanics 1:66, 1:86–87
- Bogomolny–Prasad–Sommerfeld states
definition 1:382
microscopic entropy for two-charge
systems 1:377
nonextremal black holes 1:379
- Boltzmann–Einstein formula, and
thermodynamic fluctuations 3:357
- Boltzmann equation
consequences 1:308
and hard spheres 1:308
hydrodynamics 1:309
kinetic models 2:199
overview 1:306–312
and quantum kinetic theory 1:310
solution existence 1:309
stationary problems 1:310
- Boltzmann–Grad limit, multiscale
approaches 3:470–471
- Boltzmann heuristic argument,
overview 1:306
- Bondal–Kapranov enhancements, and
derived categories 2:44
- Bondi–Metzner–Sachs group, and
conformal infinity 1:224
- Bondi–Sachs mass-loss formula, and
conformal infinity 1:225
- Boomerons, and integrable PDEs 3:116
- Boost-rotation symmetric spacetimes,
Einstein equations 2:171
- Bootstrap form-factor program
recent developments 5:339
2D QFT 5:338
- Borel's model, cohomology 2:242
- Bose–Einstein condensation
bosons 4:413
coherence 1:316
definition 1:312
in ideal gases 1:313
in interacting gases 1:314
overview 1:312–318
quantum statistics overview 1:82
superfluidity 1:316
and superfluids 5:117
- Bose–Fermi alternative
in axiomatic QFT 1:235
 N -point functions 1:236
- Bose gas, and topological
defects 5:261–262
- Boson–fermion equivalence, in 2D
model 5:331
- Bosonization–fermionization model,
overview 5:328
- Bosons
Bose–Einstein condensation 4:413
closed bosonic string field theory 5:100
degeneracy 5:116
in external fields, overview 1:318–326
in four dimensions 3:574
Goldstone bosons 5:200
many-body formalism 1:324
massless, and topological
singularities 4:224
nonlinear sigma models 4:414
one-particle formalism 1:324
open bosonic string field theory 5:95
relativistic, field algebras and quasifree
representations 1:322
superconductivity 4:414
in three dimensions 3:577
transformation, and quantum
fields 4:224
in two dimensions 3:576
variants, and random matrix theory
supersymmetry method 5:154
- Bost–Connes system, and class field
theory 3:603
- Böttcher coordinate, holomorphic
dynamics 2:656
- Bott periodicity
 C^* -algebras 3:250
in K -theory 3:247
and twisted K -theory 3:252–253
- Boundary conditions
atmospheric PEs 2:538
and currents 2:524
Dirichlet boundary condition 1:382
equilibrium statistical mechanics
overview 1:59
four-manifolds 2:462
and general relativity theory 1:610
ocean PEs 2:536
and phase transitions 4:54
and semi-classical trace formula 4:517
- Boundary conformal field theory
annulus partition function 1:336
boundary entropy 1:338
bulk–boundary operator product
expansion 1:339
extended algebras 1:339
overview 1:333–340, 1:336
stochastic Loewner evolution 1:339
- Boundary control method, wave
propagation
extensions 1:345
overview 1:340–345
- Boundary-driven models, interacting
stochastic particle systems 3:132
- Boundary entropy, in boundary conformal
field theory 1:338
- Boundary-integral methods, for
multicomponent fluid flow 3:136
- Boundary layers, and multiscale
approaches 3:474
- Boundary point-vortex motion
and Kirchhoff–Routh–Lin theory 4:75
Lin's special Green's function 4:76

- Boundary point-vortex motion (*continued*)
 method of images 4:73
- Boundary quantum groups, Yang–Baxter
 equation 1:189
- Boundary-value problem
 for integrable equations 1:346–351
 nonlinear Schrödinger equation
 1:349
- Bounded convex domains, and billiards
 caustics 1:297
 integrable billiards 1:298
 length spectrum 1:299
 overview 1:296–300
 periodic orbits 1:298
- Bounded domains, quotients, and complex
 analysis 4:550
- Bounded jumps, random walks in random
 environments 4:363
- Bounded variation
 characteristics 2:412
 convex functionals on 2:414
 function structure 2:413
 nonconvex functionals on 2:415
 and operads 3:615
 quantization in Lagrangian
 formalism 2:566
 sets of finite perimeter 2:413
- Bounding tori
 and attractors 1:485
 embeddings 1:487
- Bound states
 excitation in continuous spontaneous
 localization model 4:273
 and massive field theory 4:484
 overview 4:475–487
 in quantum electrodynamics 4:171
- Boussinesq approximations, in
 thermohydraulics 3:493
- Boussinesq equation, and bi-Hamiltonian
 formulation 1:294–295
- Boussinesq local balances, and Newtonian
 fluids 3:493
- Box dimensions, in dynamics 2:395
- BPS states *see*
 Bogomolny–Prasad–Sommerfeld states
- Bracket polynomial, in knot theory
 3:221
- Braided groups
 basic definitions 1:354
 Hopf algebras in 1:355
 and Jones polynomial 3:180
 and knot theory 3:221, 5:466
 and nonabelian Chern–Simons theory
 4:498
 overview 1:351–359
 quantum groups 2:694
- Branched manifolds, stretching and
 squeezing 1:478
- Branching patterns, and equivariant
 bifurcation theory 5:188
- Branes
 and black holes 1:373–382, 1:380
 cosmological brane worlds 1:370
 D-branes 1:175, 1:377, 2:384, 2:464,
 3:522, 5:106, 5:136
 and duality 1:375
 gauge groups on 1:360
 and gauge theory, overview 1:360–367
 and Hawking radiation 1:379
 higher-dimensional gravity 1:367
 Kaluza–Klein models 1:369
 large extra dimensions 1:368
 and near-extremal entropy 1:379
 and nonextremal black holes 1:379
 in open-string theory 3:271
 overview 1:367–372
 Randall–Sundrum brane worlds
 1:369
 in string theory 1:375
 and supergravity 5:127
- Brauer–Wall group, and Clifford
 algebras 1:524
- Breaking water waves
 governing equations 1:383
 nonlinear model equations 1:384
 overview 1:383–386
- Breathers
 and integrable PDEs 3:116
 and sine-Gordon equation 4:578
 as solitons 4:603
- Brill–Noether theory, and Riemann
 surfaces 4:426
- Brown–Feynman reduction, loop
 integrals 4:38
- Brownian motion
 fermionic Brownian motion 4:11
 and interacting stochastic particle
 systems 3:131–132
 mesoscopic theory 3:478
 microscopic theory 3:478
 strongly damped Brownian particle
 5:87
- BRST formalism
 geometric setting 1:387
 Hamiltonian BRST construction 1:389
 homological algebra results 1:386
 Koszul–Tate differential 1:388
 with Koszul–Tate differential and
 longitudinal complex 1:388
 Lagrangian BRST construction 1:390
 longitudinal complex 1:388
 and perturbative renormalization
 theory 4:41–47
 and topological sigma
 models 5:290–291
- BRST quantization
 approaches 1:390
 gauge model deformations 1:391
 gauge theory renormalization 1:391
 and graded Poisson algebra 2:564
 overview 1:386–392
 spacetime locality 1:390
 string theory applications 1:391
- BS equations
 and QFT analytic program 4:473
 and 2PI kernel 4:484
- BT *see* Bäcklund transformations
- BTZ black hole, and Chern–Simons
 theory 4:501
- Bulk–boundary operator product
 expansion, in boundary CFT 1:339
- Bulk-driven models, interacting stochastic
 particle systems 3:133
- Burgers-type equations, and Cauchy
 problem 1:446–454
- BV *see* Bounded variation
- ## C
- Cabibbo–Kobayashi–Maskawa quark
 mixing, and
 Glashow–Weinberg–Salam
 theory 2:212
- Calabi conjecture, in special Lagrangian
 geometry 1:400
- Calabi–Yau compactification, heterotic
 string 1:588
- Calabi–Yau geometry,
 characteristics 3:604
- Calabi–Yau manifolds, in special
 Lagrangian geometry 1:399
- Calabi–Yau space, and branes 1:380–381
- Calabi–Yau space, definition 1:382
- *-Calculi, definition 4:238
- Calculus types
 differential calculus 3:384, 4:237
 Malliavin calculus 3:383–389, 5:70
 operational calculus 4:636
 \hbar -pseudodifferential calculus 2:704
 spinor calculus 4:669
 on supermanifolds 5:132
 universal calculus 4:240
 universal differential calculus 4:237
 white-noise calculus 2:312
- Calculus of variations
 for microstructures 5:364
 and minimax principle
 closed geodesics 3:436
 compact energy hypersurfaces 3:437
 map class rigidity 3:436
 mountain-pass theorem 3:432
 nonlinear elliptic boundary-value
 problem 3:433
 overview 3:432–439
- Calibrated geometry,
 overview 1:398–402
- Calibrated submanifolds
 definition 1:398
 and special holonomy 1:398
- Calibrations, definition 1:398

- Calogero lattice, classical mechanics
overview 1:14
- Calogero–Moser–Sutherland systems
action-angle transforms and duality
1:409
classical nonrelativistic systems 1:405
classical relativistic systems 1:407
overview 1:403–411
quantum nonrelativistic systems 1:406
quantum relativistic systems 1:408
- Calogero–Moser systems
functional equations for 2:427
quantum C–M systems 4:123–129
and sine-Gordon equation 4:582
- Camassa–Holm equation
and bi-Hamiltonian
formulation 1:294–295
for breaking water waves 1:385
- Cameron–Martin space, and
integration-by-parts
formulas 3:415–416
- Canonical coherent states, properties 1:537
- Canonical commutation relations, and Lie
groups 3:295
- Canonical coordinates, on semisimple
Frobenius manifold 5:441
- Canonical ensemble, in Falicov–Kimball
model mathematics 2:284
- Canonical general relativity
ADM formalism 1:414
Ashtekar formalism 1:416
constraint structure 1:414
Dirac theory of constrained systems 1:412
evolution 1:416
overview 1:412–417
tetrad formalism 1:415
- Canonical Hamiltonian, and constrained
systems 1:612
- Canonical Lie groups, and Poisson
reduction 4:80
- Canonical measures, and semiclassical
trace formula 4:514
- Canonical Poisson structure, symplectic
manifold 3:315
- Canonical quantization
deformations 2:9
gerbes from 2:541
and knot invariants 3:216
- Canonical structure, Einstein gravity
classical constraints 5:455
constraint algebra 5:456
overview 5:454
- Canonical tilings, from root lattices
4:313
- Canonical transformations
in integrable dynamical
systems 3:107–108
phase space coordinates 1:4
- Cantor spectrum, and quasiperiodic
localization 3:338
- Capacity
classical, entanglement-assisted 1:418
quantum channels 1:418–423, 1:425,
4:142–144
- Capillary surfaces
compressibility, equations 1:444
corner point behavior 1:442
discontinuous dependence 1:438, 1:439
drops in wedges 1:443
existence questions 1:440, 1:441
fluids *vs.* solids 1:436
Gauss’ contribution 1:433
Laplace contribution 1:433
overview 1:431–445
stability questions 1:443
uniqueness and nonuniqueness 1:436
Young’s contribution 1:431
- Cartan domains
pseudo-Riemann symmetric
manifolds 1:509
Riemannian symmetric manifolds 1:505
- Cartan–Friedrichs formalism, and
Newton’s theory of gravitation 3:504
- Cartan–Killing form, as Lie algebra
tool 3:291
- Cartan model, and equivariant
cohomology, overview 2:242–250
- Cartan representations, Clifford
algebras 1:525
- Cartan subalgebras, and maximal tori 1:578
- Cartan–Thullen theorem, and complex
analysis 4:542
- Cartan-type Lie superalgebras,
characteristics 3:306
- CAs *see* Cellular automata
- Casimir function, and bi-Hamiltonian
formulation 1:292–294
- Casson’s invariant, in gauge theory 2:477
- A_∞ -Categories, operads 3:615
- Cat maps, quantum eigenstates 1:219
- Cauchy identity, and Plancherel
measure 4:349
- Cauchy–Kowalewska systems,
solving 5:166
- Cauchy problem
for Burgers-type equations,
overview 1:446–454
for compressible inviscid fluids 3:165
and general relativity theory 1:609
for incompressible perfect fluid 3:163
Liouville–Vlasov equation 2:199
and Minkowski space stability 5:15
- Causal boundary, for spacetime 1:329,
1:330, 1:332
- Causality
asymptotic causality 4:479
and dispersion relations 2:95
microcausality 1:235, 1:236, 2:95,
4:468
and spacetime topology 4:621
- Causal spin foams, in 4D 4:654
- Causal structure, in spacetime topology
4:617–623
- Caustics
in billiard system 1:297
nonoptical caustics 3:627
optical caustics 3:620–627
stationary phase
approximation 5:48
- CCR algebra
abnormal fluctuations 4:136
and quantum central-limit
theorems 4:132
- CCS *see* Canonical coherent states
- Celestial mechanics
overview 3:189–199
planetary $(1+n)$ -body problem 3:189
spin–orbit problem 5:24
stability problems 5:20–25
- Cellular automata
applications 1:466
for computer simulations 1:464
evolution law classifications 1:457
extensions 1:457
higher time derivatives 1:460
invertible CA 1:461
multidimensional CA 1:459
overview 1:455–467
in physics 1:465
as Turing machines 1:464
as universal constructors 1:464
vector CA 1:458
- Center of mass, and general
relativity 2:507
- Central charge, and Ads–CFT 1:180
- Central limit theorem, for counting
function 2:52
- Central manifolds
diffeomorphisms and periodic
orbits 1:468
and multiscale approaches 3:473
nonuniqueness 1:467
normal forms 1:468
overview 1:467–472
parameters 1:468
reduction 1:468
smoothness 1:467
- Central motion, classical mechanics
overview 1:9
- CERN, and electroweak theory 2:209
- CFD *see* Computational fluid dynamics
- CFL condition *see*
Courant–Friedrichs–Levy condition
- CFT *see* Conformal field theory
- Chain complexes
in cohomology theories 1:546
dual, and cohomology 1:548
homotopy group relationships 1:547
singular homology 1:546
Steenrod algebra 1:548

- Chain recurrence classes, in general dynamical systems 2:496
- Chains, and hyperbolic dynamical systems 2:724
- Channels, in quantum information theory
 - Clifford channels 2:330
 - and completely positive maps 1:473
 - instruments 1:476
 - Jamiołkowski isomorphism 1:474
 - observables 1:476
 - overview 1:472–477
 - parameter-dependent operations 1:476
 - preparations 1:476
 - Steinespring theorem 1:474
 - under symmetry 1:475
 - systems, states, algebras 1:472
- Chaotic hypothesis, and nonequilibrium statistical mechanics 3:533
- Chaotic size dependence, and short-range spin glasses 4:571
- Chaotic systems
 - anticipating synchronization 5:224
 - bounding tori 1:485
 - bounding tori embeddings 1:487
 - branched manifolds 1:478
 - and KAM 5:30
 - orbit basis sets 1:483
 - and strange attractors, overview 1:477–488
 - and vortex dynamics 5:394
- Chaotic system synchronization
 - applications 5:225
 - bidirectional identical synchronization 5:218
 - generalized synchronization 5:221
 - identical synchronization 5:215
 - lag synchronization 5:223
 - overview 5:213–226
 - phase synchronization 5:222
 - projective synchronization 5:224
 - spatio-temporal synchronization 5:224
 - stability 5:214
 - unidirectional identical synchronization 5:217
- Characteristic classes
 - Chern character 1:494
 - Chern–Weyl theory 1:494
 - connections 1:493
 - curvature 2-form 1:493
 - de Rham cohomology 1:492
 - Euler form 1:495
 - examples 1:494
 - formulas summary 1:495
 - overview 1:488–496
 - principal bundles 1:495
 - projections 1:493
 - unitary connections 1:493
 - vector bundles 1:488
- Characters, and compact groups 1:585
- Charge conservation, electromagnetism overview 1:47
- Charged fluids, Einstein equations 2:199
- Charge distributions, electromagnetism overview 1:43
- Charged particles, and Minkowski spacetime 1:105
- Charges, electromagnetism overview 1:42
- Charlier polynomials, characteristics 3:643
- Checkerboard generalizations, and Yang–Baxter equations 5:468
- Checkerboard IRF model, Yang–Baxter equations 5:469
- Checkerboard IRF-vertex model, Yang–Baxter equations 5:469
- Checkerboard model, square weight 5:470
- Checkerboard vertex model map to 5:470
- Yang–Baxter equations 5:469
- Chekanov’s formula, in caustics 3:626
- CH equations
 - and compactons 4:16
 - dispersionless, reduction 4:17
 - integrable peakon dynamics 4:15
 - peakon solution generalization 4:16, 4:17
- Chern character
 - noncommutative tori 3:525
 - overview 1:494
 - and twisted K -theory 3:251
- Chern classes, vector bundles 1:492
- Chern forms, index theorems 3:27
- Chern–Simons theory
 - fermion CS theory 2:407
 - functional integrals 1:497
 - and large- N duality 3:266
 - and large- N expansion 3:265
 - manifold invariants from 4:499
 - and nonabelian BF theories 1:263
 - overview 1:496–500, 3:270
 - and 3D gravity 4:501
 - as TQFT example 5:284
 - various perspectives 5:288
- Chern–Weyl theory, characteristic classes 1:494
- Chevalley groups, q -special function applications 4:116
- Chevalley theorem, and Clifford algebras 1:524
- Chiral anomalies
 - abelian and nonabelian types 1:205
 - applications 1:207
 - characteristics 1:205
 - and gauge theories 1:207
 - global structure 1:208
 - instanton physics applications 1:207
 - matching conditions 1:208
 - neutral pion decay application 1:207
 - nonrenormalization 1:206
 - theta vacuum applications 1:207
- Chiral fields
 - and conformal field theory 5:319
 - and 2D conformal models 5:335
- Chiral perturbation theory, and standard model 2:145
- Chiral symmetry, quantum chromodynamics 4:147
- Chiral theories, and conformal quantum fields 5:334
- Chord diagrams
 - algebras 3:232
 - and finite-type invariants 2:341
 - Hopf algebra structure 3:232
- Chorin–Temam methods, in fluid mechanics 2:373
- Chow groups, and intersection theory 3:155
- Chronological functions, in massive field theory 4:478
- Circle, homeomorphisms and diffeomorphisms 2:665–672
- Claspers, and finite-type invariants 2:345, 2:349
- Class field theory, and quantum statistical mechanics 3:603
- Classical action, nonabelian BF theories 1:258
- Classical channels, overview 1:475
- Classical data compression
 - asymptotic equipartition property 4:610
 - entropy and source coding 4:609
- Classical differential geometry, and integrable systems 3:78
- Classical dynamics, and semiclassical trace formula
 - canonical measures and symplectic reduction 4:514
 - Newton flows 4:513
 - nondegeneracy 4:514
 - periodic orbits 4:513
- Classical elliptic complexes
 - de Rham complex 3:24
 - Dolbeault complex 3:25
 - Lefschetz formula 3:31
 - relating complexes 3:26
 - signature complex 3:25
 - spin complex 3:26
 - twisted signature complex 3:25
 - Yang–Mills complex 3:25
- Classical fields, and Anderson–Higgs–Kibble mechanism 4:227
- Classical gauge theories, and Seiberg dualities 4:506
- Classical gravitational many-body problems
 - collisional equilibrium 2:580
 - collisional evolution 2:581
 - distribution function evolution 2:579
 - five-body problems 2:579
 - four-body problems 2:579

- general three-body problem 2:577
- many-body problems 2:579
- numerical methods 2:581
- overview 2:575–582
- restricted three-body problem 2:576
- two-body problem 2:576
- Classical groups
 - isomorphisms in small dimensions 1:503
 - overview 1:500–511
- Classical information theory, and quantum information theory 4:142
- Classical integrable systems
 - affine Toda theories 3:51
 - and field theory 3:50
 - sinh–sine-Gordon theory 3:51
- Classical knot polynomials, and finite-type invariants 2:341
- Classical Lefschetz formula, characteristics 3:31
- Classical Lie superalgebras
 - characteristics 3:306
 - overview 3:307
 - structure 3:307
- Classical Lorentz electron gas, and multiscale approaches 3:467
- Classical master equation, and Batalin–Vilkovisky quantization 1:249
- Classical mechanics
 - Coriolis forces 1:32
 - diffusion in phase space 1:24
 - equations of motion 1:3
 - evolution operators 2:707
 - and \hbar -Fourier integral operators 2:707
 - generic nonintegrability 1:15
 - integrability 1:7
 - inverse problem overview 3:156–160
 - KAM resummation scheme 1:30
 - KAM stability 1:20
 - Kepler’s laws 1:10
 - Lindstedt and Birkhoff series 1:18
 - Lorentz forces 1:32
 - multidimensional quadratures 1:9
 - Newtonian potential 1:10
 - overview 1:1–33
 - particle constraints 1:2
 - perturbing functions 1:16
 - phase space coordinates 1:4
 - and \hbar -pseudodifferential operators 2:701
 - quadratures 1:6, 1:13
 - quasiperiodicity 1:7, 1:20
 - recursion operators overview 4:371–376
 - resonances and Lindstedt series 1:23
 - resonances and stability 1:21
 - rigid body 1:11
 - singularities 1:28
 - three-body problem 1:27
- Classical moduli space, nonabelian BF theories 1:258
- Classical nonabelian BF theories
 - uses and applications 1:262
- and Yang–Mills theory 1:262
- Classical nonrelativistic
 - Calogero–Moser–Sutherland systems, overview 1:405
- Classical optics, Kramers–Krönig dispersion relation 2:88–89
- Classical orthogonal polynomials, overview 3:641
- Classical physics
 - electromagnetic fields 1:191
 - quantum free gas problem 4:302
- Classical relativistic
 - Calogero–Moser–Sutherland systems, overview 1:407
- Classical r -matrices
 - overview 1:511–518
 - and special Lie bialgebra classes 1:513
- Classical Stein manifolds, overview 1:502
- Classical supersymmetry, and supersymmetric oscillator 5:147
- Classical systems, reflection positivity and phase transitions
 - Gaussian domination 4:380
 - infrared estimates 4:381
 - main estimate 4:379
 - overview 4:377
- Classical weak measurement
 - overview 4:277
 - postselection 4:277
 - time-continuous measurement 4:278
- Classical Yang–Baxter equations
 - characteristics 5:471
 - and quantum inverse-scattering method 5:468
- Clebsch action, in Hamiltonian fluid dynamics 2:597
- Clifford algebras
 - adjacent dimension spaces 1:524
 - and Brauer–Wall group 1:524
 - Cartan representations 1:525
 - change conjugation 1:526
 - and Chevalley theorem 1:524
 - complexification 1:524
 - definitions 1:520
 - Dirac representations 1:525
 - graded types 1:522
 - grading 1:523
 - Hermitian scalar products 1:527
 - inner product spaces 1:519
 - Majorana spinors 1:526
 - multivector representations 1:527
 - notation 1:519
 - overview 1:518–530
 - Pauli representations 1:525
 - pin groups 1:529
 - Radon–Hurwitz numbers 1:527
 - representation construction 1:527
 - simplicity theorem 1:525
 - spin groups 1:529
 - spinor groups 1:528
- structure 1:524
- universal property 1:523
- vector fields 1:528
- vector space structure 1:523
- volume element and centers 1:524
- Weyl representations 1:525
- Clifford channels
 - and finite Weyl systems 2:330
 - and graph representations 2:332
 - and stabilizer formalism 2:331–332
- Cloning, quantum states 3:628–632
- Closed bosonic string field theory, overview 5:100
- Closed geodesics, on compact manifolds 3:436
- Closed manifolds, CMC data 2:177
- Closed orbits, and semi-classical spectra
 - bifurcations 4:517
 - canonical measures and symplectic reduction 4:514
 - case with boundary 4:517
 - Feynman integral derivation 4:516
 - integrable case 4:516
 - inverse spectra problem 4:517
 - maximally degenerated case 4:517
 - Newton flows 4:513
 - nondegeneracy 4:514
 - overview 4:512–518
 - periodic orbits 4:513
 - Selberg trace formula 4:513
 - semiclassical Schrödinger operators on Riemannian manifolds 4:513
 - smoothed density of states 4:515
 - spectral densities 4:515
 - Weyl expansion 4:516
- Closed string theory
 - in open bosonic string field theory 5:100
 - overview 3:271
- Cluster expansions, for lattice model
 - abstract polymer models 1:532
 - correlations 1:534
 - high-temperature expansions 1:534
 - low-temperature expansions 1:535
 - Mayer expansion 1:531
 - overview 1:531–537, 1:533
- Clusters
 - N -point functions 1:237
 - in perturbation theory 2:58
- CMC *see* Constant mean curvature
- CMS systems *see*
 - Calogero–Moser–Sutherland systems
- CM systems *see* Calogero–Moser systems
- CNOT *see* Controlled-NOT
- Coalescence catastrophe
 - from long-range attraction 1:85
 - from short-distance attraction 1:84
- Coarse-grainings
 - and functional integration 2:446
 - and multiscale approaches 3:473
 - and phase transitions 4:55

- Cobordism theorem, overview 1:552
- Codifferential operator, in
 - electrodynamics 1:142
- Codimension, in optical caustics 3:626
- Coding theorems, and quantum channel capacity 1:428
- Coexisting regimes, and multiscale approaches 3:469
- Cogravity, and bicrossproduct Hopf algebras 1:265
- Coherent states
 - applications 1:544
 - Bose–Einstein condensates 1:316
 - examples 1:539
 - general definition 1:538
 - overview 1:537–545
- Coherent structures, from turbulent flows 5:414
- Cohomological field theory, and
 - Mathai–Quillen formalism and supersymmetric quantum mechanics 3:397
 - and topological sigma model 3:397
- Cohomology rings, symplectic quotients 2:605
- Cohomology theories
 - chain complexes and homology 1:546
 - classifying spaces 1:551
 - compact complex manifolds 4:557
 - differential geometry overview 1:39
 - and dual chain complexes 1:548
 - Hodge decomposition 1:550
 - K -theory 1:552
 - Lie groups 1:551
 - moduli spaces 1:551
 - overview 1:545–553
 - projective spaces 1:551
 - simplicial homology 1:546
 - singular cohomology 1:548
 - smooth manifolds 1:549
- Coisotropic submanifolds
 - and groupoids 3:315
 - Poisson reduction 4:81
- Collet–Eckmann condition, and stochasticity 5:348
- Collisional effects, classical gravitational many-body problems 2:580
- Color confinement, mechanisms 3:280
- Colored enumeration, Redfield–Pólya theory 1:558
- Combescure transformations, multidimensional quadrilateral lattices 3:82
- Combinatorial classification, subfactors 5:387
- Combinatorial construction, Kontsevich integral
 - computation recipe 3:236
 - Drinfeld associator and rationality 3:237
 - example 3:237
 - knot, sliced presentation 3:236
- Combinatorial equivalence, holomorphic dynamics 2:656
- Combinatorial problems, and statistical mechanics
 - constraint satisfaction problems 5:51
 - counting problem complexity 5:53
 - message-passing algorithms 5:53
 - overview 5:50–55
 - satisfiability and spin glass models 5:52
- Combinatorics
 - alternating-sign matrices 1:572
 - asymptotic information from generating functions 1:560
 - basic terminology 1:554
 - binomial sums 1:573
 - fully packed loop configurations 1:570
 - generating functions 1:556
 - generating function solutions 1:559
 - heights 2:62
 - hypergeometric series 1:573
 - lattice paths 1:565
 - and limit laws F_β 3:103
 - local statistics 2:62
 - nonintersecting paths 1:569
 - overview 1:553–576
 - partition function 2:62
 - perfect matchings 1:568
 - plane partitions 1:570
 - Redfield–Pólya theory 1:558
 - rhombus tilings 1:570
 - six-vertex model 1:572
 - theory of heaps 1:562
 - tilings 1:568
 - transfer matrix method 1:564
 - vicious walkers 1:570
- Communication, chaos synchronization applications 5:225
- Commutative C^* -algebras
 - positive maps in 4:89
 - states on 4:90
- Commutativity, Bäcklund transformations 3:117–118
- Commutators, for finite-type invariant fundamental theorem 2:345
- Commuting linear maps, and symmetry 5:184
- Commuting transfer matrices, eight-vertex and hard-hexagon models 2:158
- Comoving frame, classical mechanics overview 1:11–12
- Compacification, definition 1:382
- Compact classical homogeneous manifolds, overview 1:503
- Compact complex manifolds
 - biholomorphic classification 4:554
 - birational classification 4:551
 - deformations, and Kodaira–Spencer algebra 2:22
 - homogeneous types 4:558
 - overview 4:551–559
 - topology, invariants, cohomology 4:557
- Compact energy hypersurfaces, closed characteristics 3:437
- Compact groups
 - basic notions 1:580
 - characters 1:585
 - explicit representation construction 1:584
 - and Fourier transform
 - Haar measure and invariant integral 1:582
 - noncommutative Fourier transform 1:583
 - orthogonality relations 1:582
 - Peter–Weyl theorem 1:582
 - main theorems 1:581
 - overview 1:576–586
 - representation examples 1:581
 - tensor product multiplicities 1:585
 - unirreps classification 1:584
 - weight decomposition 1:584
 - Weyl character formula 1:585
- Compactification, superstring theory
 - extended supersymmetric 1:590
 - higher-dimensional 1:590
 - overview 1:586–595
 - and string theory 1:588, 5:103
 - and supergravity 1:590
- Compact Lie groups
 - complexification 1:577
 - and cotangent bundle reduction 1:663
 - examples 1:577
 - and maximal tori 1:578
 - and semisimple Lie algebras 1:577
 - simple, classification 1:578
 - Weyl group 1:580
- Compact manifolds, closed geodesics on 3:436
- Compactness, in space BV 2:412
- Compact objects, and gravitational waves 2:584
- Compactons, and CH equations 4:16
- Compact operators, and spectral theory 4:638
- Compact Riemannian manifolds, sharp Sobolev inequalities on 3:36
- Compact Riemann surfaces
 - moduli 4:427
 - properties 4:421
- Compact special Lagrangian manifolds
 - deformations 1:400
 - obstructions 1:401
- Comparative entanglement
 - asymptotic conversion 2:234
 - protocols 2:233
 - pure state conversion and majorization 2:234

- Complete decoupling, in effective field theory 2:139–140
- Completely positive maps
 C^* -algebras 4:91
 and channels 1:473
 and continuity 4:92
 in mathematical physics 4:92
 and matrix algebras 4:92
 and quantum channels 1:473
 and quantum dynamical semigroups 4:159–160
 Steinespring representation 4:93
- Completeness
 and nilpotent Lie groups 4:96
 overview 4:475–487
 and QFT analytic program 4:473
 and scattering theory 4:464
- Complex analysis
 and algebraic geometry 4:548
 compact complex manifolds 4:551–559
 domains in C^n 4:541
 mapping theory 4:547
 Montel's theorem and Fredholm mappings 4:546
 overview 4:540–550
 Stein manifold overview 4:543
- Complex angular momentum, and analyticity 4:472
- Complex classical groups, overview 1:501
- Complex geometry, and contact manifolds 1:634
- Complexification
 Clifford algebras 1:524
 compact Lie group 1:577
 and spectral theory 4:637
- Complexity, counting problems 5:53
- Complex multiplication system, and class field theory 3:604
- Complex renormalization
 and little Mandelbrot sets 5:346
 and polynomial-like maps 5:346
- Complex space, deformation 2:18
- Complex spin groups, Clifford algebras 1:530
- Complex structures, Clifford algebras 1:519
- Composite-fermion theory
 and fractional quantum Hall effect 2:410
 overview 2:406
 physical interpretation 2:406
- Composition law, and quantum dynamical semigroups 4:159–160
- Compressibility, in capillary surfaces, equations 1:444
- Compressible flows
 inviscid compressible fluid flows 1:596
 overview 1:595–603
 viscous compressible fluid flows 1:602
- Compressible fluids, and symmetrization 5:165
- Compressible inviscid fluids, Cauchy problem 3:165
- Computational fluid dynamics, for turbulence theories 5:298
- Computer simulations, with cellular automata 1:464
- Concurrence, properties 2:235–236
- Condensation
 Bardeen–Cooper–Schrieffer–Leggett pair condensation 5:118
 Bose–Einstein condensation 1:82, 1:312–318, 4:413, 5:117
 droplet condensation 3:262
 local response 5:119
 tachyon condensation 3:522
- Condensed matter
 bosons 4:413
 fermion overview 4:411
 and renormalization group 4:407
- Conditional expectation, in quantum relative entropy 4:179
- Cone spaces, singular reduction 5:196
- Configuration space integrals, for finite-type invariant fundamental theorem 2:344
- Confinement, and quantum chromodynamics 4:148
- Confluence, and isomonodromy 3:174
- Confluent hypergeometry, characteristics 3:640
- Conformal blocks, and Riemann surfaces 4:428
- Conformal covariance, and tochastic Loewner evolutions 5:85
- Conformal embeddings, and spacetime boundaries 1:328
- Conformal field theory
 and absolute Galois group 3:604
 anti-de Sitter spacetime–CFT correspondence 1:174–183
 and braided monoidal categories 1:352
 and dilogarithm 3:601
 and fractional quantum Hall effect 2:408
 general structure 5:318
 and integrable QFT 3:57
 and modular invariance 5:321
 modular invariance 1:335
 and open bosonic string field theory 5:95
 and operads 3:613
 operator product expansion 3:619
 and quantum geometry 5:139
 and stochastic Loewner evolutions 5:83
 stress tensor and Ward identities 1:333
 superconformal field theory 3:439–448
 Virasoro algebra 1:334
 and von Neumann algebras 5:389
- Conformal flow
 definition 2:516
 general relativity 2:516
- Conformal infinity
 asymptotically flat spacetimes 1:223
 BMS group 1:224
 Bondi–Sachs mass-loss formula 1:225
 Minkowski spacetime 1:222
 overview 1:221–226
 Peeling property 1:224
- Conformal method
 for Einstein equation 2:176
 and holomorphic dynamics 2:659
- Conformal quantum field theory, and chiral theories 5:334
- Conformal symmetry group, overview 5:317
- Conformal thin sandwich, and Einstein constraint equations 2:178
- Conformal volume collapse, and Bianchi model reduction 2:618
- Conjugate representations, Clifford algebra 1:526
- Conjugation properties, in semisimple abelian modular categories 1:358
- Conley's theory, in general dynamical systems 2:496
- Connections
 covariant derivative 4:241
 differential geometry overview 1:37
 noncommutative tori 3:525
 in quantum principal bundles 4:241
- Connectivity, and holomorphic dynamics 2:658
- Conservation laws
 baryon–antibaryon asymmetry 5:170
 baryon and lepton numbers 5:169
 discrete symmetries 5:169
 gauge theories 5:167
 general relativity 5:171
 and integrable PDEs 3:118
 Liouville–Vlasov equation 2:199
 overview 5:166–172
 spontaneous symmetry breaking 5:168
 supersymmetry 5:171
 in symmetrization 5:161
 topological changes 5:170
- Conservation of mass, in fluid mechanics 2:365
- Conservativity, and quantum dynamical semigroups 4:159–160
- Conserved systems, spinoidal decomposition 4:49
- Conserved vector current, and electroweak theory 2:209
- Consistency conditions, wave equations 4:395
- Consistent anomalies, covariant anomaly comparison 1:210

- Constant mean curvature
 - on closed manifolds 2:177
 - and minimal surfaces 3:430
- Constants of motion
 - incompressible perfect fluid 3:161
 - and integrable many-body problems 3:109
- Constant term identities, and root systems 4:112
- Constrained systems
 - Dirac algorithm 1:612
 - Dirac theory 1:412
 - first- and second-class constraints 1:613
 - general structure 1:414
 - geometric description 1:616
 - overview 1:611–616
 - quantization 1:615
- Constraint equations
 - and conformal thin sandwich 2:178
 - and general relativity theory 1:608
 - gluing solutions 2:180
 - quasispherical ansatz 2:179
 - and space + time field decomposition 2:174
- Constraint satisfaction problems, ombinatorics 5:51
- Constraints formula, and Poisson reduction 4:82
- Constructive quantum field theory
 - asymptotic freedom 1:623
 - effective potentials 1:625
 - Euclidean quantum fields 1:617
 - infrared behavior 1:629
 - large fields 1:627
 - overview 1:617–631
 - perturbation theory 1:619
 - power counting 1:621
 - regularized free field 1:618
 - small fields 1:626
 - ultraviolet limit 1:629
- Contact manifolds
 - and complex geometry 1:634
 - definitions and examples 1:631
 - existence and classification 1:633
 - and geometric optics 1:635
 - Hamiltonian dynamics 1:635
 - local theory 1:632
 - overview 1:631–636
 - Reeb vector field 1:634
 - Riemannian geometry 1:634
 - and symplectic geometry 1:633
- Contact processes, and percolation 4:27
- Continuity
 - in axiomatic QFT 1:235
 - in basic convex analysis 1:642
 - C^* -algebras 3:250
 - electromagnetism overview 1:47
 - and entanglement measures 2:235
 - and measures on loop space 3:416
 - and quantum dynamical semigroups 4:159–160
- Continuous Hahn polynomials, characteristics 3:644
- Continuously varying landscapes, stochastic resonance 5:91
- Continuous semigroups, exact renormalization group as 2:277
- Continuous spontaneous localization model
 - basics 4:271
 - bound state excitation 4:273
 - consequences 4:272
 - interference 4:272
 - overview 4:270
 - random walk 4:273
- Continuous symmetries, equilibrium statistical mechanics 1:65
- Continuous systems, phase transitions
 - boundary condition dependence 4:54
 - coarse graining transformations 4:55
 - competing species 4:55
 - continuum model 4:57
 - contours and indicators 4:58
 - Ising model 4:55
 - Kac potentials 4:57
 - overview 4:53–59
 - and statistical mechanics 4:53
 - van der Waals liquid–vapor transition 4:56
- Continuous-time random walks in random environments, characteristics 4:366
- Continuous q -ultraspherical polynomials
 - characteristics 4:110
 - Macdonald polynomials 4:114
- Continuous wavelet transform
 - analysis 5:426
 - examples 5:427
 - higher dimensions 5:428
 - synthesis 5:427
- Contour dynamics, in vortex dynamics 5:397
- Contractions, Clifford algebras 1:520
- Controllability
 - as control problem 1:636
 - ordinary differential equations 1:638
 - partial differential equations via Coron return method 1:641
- Controlled-NOT, and fault-tolerant gates 4:199
- Control problems
 - overview 1:636–642
 - partial differential equations 1:640
 - Schrödinger equation 1:641
- Convex analysis
 - basic analysis 1:642
 - continuity 1:642
 - in duality 1:644
 - lower-semicontinuity 1:642
 - Moreau–Fenchel conjugate 1:643
 - overview 1:642–652
- Convex entropy, and symmetrization 5:161
- Convex functional, on bounded variation 2:414
- Convexity
 - and entanglement measures 2:235
 - and Hamiltonian group actions 2:603
- Convex potentials, in transportation maps 3:633
- Convex programming, convex variational problems in duality 1:647
- Cooper pairs, quantum statistics
 - overview 1:83
- Coordinate space, and renormalization 4:402
- Copenhagen Interpretation, quantum mechanics 4:265
- Corank, in optical caustics 3:626
- Coriolis forces, classical mechanics
 - overview 1:32
- Corner transfer matrices, eight-vertex and hard-hexagon models 2:162
- Coron, return method, controllability 1:641
- Correlated states, finitely *see* Finitely correlated states
- Correlation functions
 - Plancherel measure 4:349
 - in random matrix theory 4:340
- Cosmic censorship, and gravitational collapse 1:672
- Cosmic shear, and gravitational lensing 2:575
- Cosmological brane worlds, characteristics 1:370
- Cosmological dynamics, brane worlds 1:371
- Cosmological models, Einstein equation solutions 2:172
- Cosmology
 - Einstein–Cartan theory of gravitation 2:194
 - mathematical cosmology 1:653–658
 - Newtonian cosmology 3:507
 - QFT applications 4:214
 - quantum cosmology 4:153–158
 - and string theory phenomenology 5:110
- Cost functions
 - duality in mass transport 1:648
 - minimization over functional space 3:4
- Cosymplectic submanifolds, and Poisson reduction 4:82
- Cotangent bundle reduction
 - amended potential 1:660
 - connections 1:659
 - dynamics reconstruction 1:661
 - embedding version 1:659
 - fibration version 1:660

- free rigid body reconstruction phases 1:664
 - gauged Poisson structures 1:665
 - heavy top reconstruction phases 1:664
 - Kaluza–Klein construction 1:660
 - magnetic terms and curvature 1:660
 - mechanical connection 1:660
 - overview 1:658–667
 - simple mechanical system reconstruction symmetry 1:663
 - Sternberg and Weinstein space symplectic leaves 1:666
 - Cotangent bundles
 - in classical mechanics 4:373
 - Lie groupoids 3:316
 - Liouville form 3:314
 - Poisson manifold 3:317
 - shifted 2:564
 - star products on 2:32
 - and symplectic reduction 5:192
 - Counter-term, formulation 4:403
 - Counting problems
 - asymptotic counting formulas 4:327
 - central limit theorem for 2:52
 - complexity 5:53
 - power counting 1:621
 - Coupled dimer antiferromagnet, and quantum phase transitions 4:291
 - Coupled level-set volume-of-fluid algorithms, for multicomponent fluid flow 3:140
 - Coupling constants, in effective field theory 2:140
 - Couplings, classical mechanics overview 1:17
 - Courant–Friedrichs–Levy condition, in fluid mechanics 2:372
 - Covariance, in axiomatic QFT 1:236
 - Covariant anomalies, consistent anomaly comparison 1:210
 - Covariant cloning maps, and optimal cloning 3:629
 - Covariant derivatives
 - and A–B geometry 1:196
 - strong connections 4:241
 - Covariant functors, QFT as 1:200
 - Covariant path integral, spin foams from 4:650
 - CP maps *see* Completely positive maps
 - CP violation, and
 - Glashow–Weinberg–Salam theory 2:212
 - Critical coupling, abelian Higgs vortices 1:156
 - Critical dynamics, and interacting stochastic particle systems 3:134
 - Critical fields
 - and Ginzburg–Landau model 5:357
 - and renormalization groups 2:276
 - Critically self-similarity, and gravitational collapse 1:670
 - Critical phenomena, in gravitational collapse
 - dynamical systems picture 1:669
 - overview 1:668–673
 - statistical mechanical analogy 1:671
 - Critical points
 - equilibrium statistical mechanics overview 1:76
 - finite group symmetry breaking 2:325
 - percolation theory 4:23
 - Critical surface mass density, in gravitational lensing 2:569
 - Crossing manifolds, and dispersion relations 2:97
 - Crossovers
 - in random matrix theory 4:343
 - and symmetry breaking 5:203
 - Cryptography, quantum, breaking 4:269
 - Crystalline surfaces, and dimer model scaling limits 2:66
 - Crystals
 - equilibrium crystals 5:61
 - insulating, heat conduction in 2:376
 - in Ising model 5:463
 - liquid crystals 3:320–328
 - nematic liquid crystals 5:262
 - periodic photonic crystals 3:491
 - polymeric liquid crystals 3:565
 - CSD *see* Chaotic size dependence
 - CSL *see* Continuous spontaneous localization model
 - CSPs *see* Constraint satisfaction problems
 - CSS *see* Critically self-similarity
 - CSS codes, and quantum error correction 4:198
 - CTMs *see* Corner transfer matrices
 - CTS *see* Conformal thin sandwich
 - Cubic lattice, and percolation theory 4:22
 - Cuntz C^* -algebras, characteristics 1:397
 - Cup product
 - C^* -algebras 3:250
 - in K -theory 3:247
 - and twisted K -theory 3:252–253
 - Current algebras
 - and groups 1:677
 - for hadrons 1:674
 - nonrelativistic 1:677
 - overview 1:674–679
 - Ringel–Hall algebras 2:317
 - Sugawara, Kac–Moody, Virasoro 1:676
 - Current conversion, differential forms 1:144
 - Currents
 - and Anderson–Higgs–Kibble mechanism 4:227
 - and boundary and mass 2:524
 - with finite mass 2:525
 - persistent, Ginzburg–Landau equations 2:549
 - and plateau problem 2:526
 - rectifiable currents 2:525
 - solutions and tools 2:527
 - Curvature
 - in cotangent bundle reduction 1:660
 - differential geometry overview 1:38
 - Einstein–Cartan theory 2:191
 - Curvature 2-form, and characteristic classes 1:493
 - Curved spacetimes
 - Dirac equation 2:69
 - gravity as 2:489
 - and QFT 4:202–212
 - relativistic wave equations 4:395
 - Curves, differential geometry overview 1:33
 - Cutoff regularization, perturbation theory 4:32
 - C-valued function, and Schrödinger equation 1:318–319
 - CVC *see* Conserved vector current
 - Cycle map, and intersection theory 3:153
 - Cyclic systems, in Levi–Civita 4:373
 - Cyclotomic expansion, Habiro ring 2:354
 - Cylindrical waves, and Einstein equation solutions 2:170
- D**
- D’Alembert’s paradox, and incompressible perfect fluid 3:162
 - Damped harmonic oscillator
 - characteristics 4:161
 - and dispersion relations 2:90
 - Darboux–Bäcklund transformations, infinite Toda lattice 5:238
 - Darboux system, discrete geometry 3:80
 - Darboux-type transformations, multidimensional quadrilateral lattices 3:82
 - Data compression
 - classical data compression 4:609
 - quantum data compression 4:611
 - Data dependence, and viscous incompressible fluids 5:370
 - Davey–Stewartson-II, and delta-bar problem method 2:38
 - D-branes *see* Dirichlet branes
 - $D1$ – $D5$ – N system, and five-dimensional black holes 1:375
 - Debye’s conjecture, and heat transport 3:549
 - Decay, wavelet coefficients 5:434
 - Decomposition
 - block matrix decomposition 3:299
 - hydrodynamics 3:362
 - Lie algebras 3:292
 - and multiscale approaches 3:471

- Decomposition (*continued*)
 and wavelet denoising 5:436
- Deconfinement, and quantum chromodynamics 3:279
- Decoupling, slow and fast motions 1:228
- Deformation, algebraic intersection products 3:152
- q -Deformation coordinate algebras, and quantum groups 2:692
- q -Deformation enveloping algebras, and quantum groups 2:690
- Deformation quantization
 conventional quantization 2:6
 equations 2:4
 and operads 3:614
 overview 2:1–9, 2:24
 Poisson manifolds 2:23
 and QFT 2:7
 and representation theory, overview 2:9–16
 simple harmonic oscillator 2:5
 star products 2:2
- Deformations
 compact special Lagrangian m -folds 1:400
 gauge model deformations 1:391
 H_0 3:213
 and integrable systems 3:68
 and knot homologies 3:215
 knot homologies 3:209
 Poisson bracket on symplectic manifold
 Berezin-type star products 2:31
 definitions 2:25
 groups actions on star products 2:29
 overview 2:24–34
 star product classification 2:27
 star product construction 2:26
 star products on cotangent bundles 2:32
 star products and function subclasses 2:30
- Deformation theory
 complex space 2:18
 definitions and examples 2:16
 and differential graded Lie algebras 2:21
 first-order, algebras 2:18
 formal, algebras 2:19
 formal quantization of manifolds 2:19
 and Grothendieck-type scheme 2:18
 and holonomic quantum fields 2:661
 Kodaira–Spencer algebra and manifolds 2:22
 matrix families as 2:18
 and Maurer–Cartan equation 2:22
 overview 2:16–24
 and quantized universal enveloping algebras 2:19
 quantum plane 2:19
 ringed space products 2:17
 Schlesinger’s theorem 2:20
 singularities 2:18
 versal deformations 2:20
- Deformed BF theory, uses and applications 1:263
- Deformed flat connection, Frobenius manifolds 5:439
- Degeneracy
 bosons and fermions 5:116
 and Pirogov–Sinai theory 4:60
- Degeneracy loci, as intersection theory application 3:154
- Degenerate Hessians, and functional integration 2:442
- Deligne–Simpson problem, and Riemann–Hilbert problem 4:440
- Delta-bar approach, integrable systems
 data reductions 2:39
 Davey–Stewartson-II and Ishimori-I equations 2:38
 discrete systems 2:38
 explicit solutions 2:39
 KP-II hierarchy 2:36
 multidimensional problems 2:38
 nonlocal and local delta-bar problem 2:35
 nonsingular solutions 2:39
 Novikov–Veselov hierarchy 2:37
 overview 2:34–41
 quasiclassical limit 2:40
 Riemann surfaces 2:40
- Delta-bar dressing method, and multidimensional integrable lattices 3:80
- Delzant’s theorem, and toric manifolds 2:604
- Denjoy theory, and circle diffeomorphisms 2:670
- Denoising
 and image processing mathematics 3:3
 wavelet approach 3:9
 wavelets applications 5:424
- Density
 equilibrium statistical mechanics overview 1:67
 Newtonian fluids, state equations 3:493
 semi-classical spectra 4:515
 in space BV 2:412
- Density matrix
 C^* -algebras 4:90
 quantum statistics overview 1:82–83
- De Rham cohomology
 and characteristic classes 1:492
 overview 1:550
- De Rham elliptic complex, characteristics 3:24
- q -Derivative, in q -hypergeometric series 4:107
- Derived categories
 Bondal–Kapranov enhancements 2:44
 and Landau–Ginzburg models 2:44
 in open-string B model 2:41
 overview 2:41–46
 and π -stability 2:45
- Derived functors, and spectral sequences 4:629
- Derrida’s models, spin glasses 3:408
- Descent equations, and anomalies 1:210
- De Sitter spacetimes, and Einstein equations solutions 2:166
- Desynchronization motion, bidirectional identical synchronization 5:221
- Deterministic chaos, isochronous systems 3:170
- Determinantal correlation functions, in random matrix theory 2:586
- Determinantal random fields
 counting function central limit theorem 2:52
 ergodic properties 2:51
 Gibbsian properties 2:52
 mathematical framework 2:47
 overview 2:47–54
- Determinantal random point fields
 discrete translation-invariant point fields 2:50
 Fermion gas 2:49
 immanantal processes 2:53
 Markov process nonintersecting paths 2:51
 measures on partitions 2:50
 Pfaffian processes 2:53
 random matrix models 2:49
- Determinants
 and abelian BF theories 1:260
 as Gaussian integrals 5:153
 Green’s functions from 5:153
- Deterministic chaos, in integrable dynamical systems 3:107
- Devetak’s theorem, and quantum channel capacity 1:428
- Deviation theory
 geodesic deviation 2:491
 large deviation theory 3:261–263
- DGA *see* Differential graded algebras
- DHR representations, and algebraic QFTs 1:203
- Diagrammatic techniques, in perturbation theory
 examples and applications 2:59
 graphs and trees 2:56
 historical excursus 2:55
 multiscale analysis 2:58
 overview 2:54–60
 paradigmatic example 2:55
- Diffeomorphisms
 central manifolds 1:468
 circle
 Denjoy example 2:670
 overview 2:665–672

- Poincaré Denjoy theorem 2:670
- Poincaré rotation number 2:669
- C^1 -generic diffeomorphisms 2:501
- and Hamiltonian system stability 2:632
- homoclinic orbits in 2:673
- and Hopf algebras 2:684
- as hyperbolic dynamical system
 - example 2:724
- as infinite-dimensional Lie group
 - example 3:41
- and knot invariants 3:217
- nonabelian BF theories 1:258
- wild behavior, local C^2 -genericity 2:498
- q -Difference equation
 - and Askey–Wilson polynomials 4:110
 - big q -Jacobi polynomial 4:111
 - continuous q -ultraspherical polynomials 4:111
 - in q -gauss hypergeometric series 4:108
 - Macdonald polynomials 4:113
- Difference of two Diracs, and turbulence 5:412
- Differentiable manifolds, differential geometry overview 1:33
- Differentiable structures, for manifolds 2:386
- Differential algebras, and equivariant cohomology 2:243
- Differential calculus
 - on algebra 4:237
 - on Wiener space 3:384
- Differential equations
 - elliptic 2:216–228
 - ODEs 1:638
 - and separation of variables
 - eigenvalue problem separation 4:533
 - Fourier method 4:526
 - Hamilton–Jacobi equation on Riemannian manifolds 4:527
 - Liouville integrable systems 4:534
 - overview 4:526–535
 - PDEs in general setting 4:527
 - stochastic 5:63–71
- Differential forms
 - and electrodynamics
 - codifferential operator 1:142
 - current conversion 1:144
 - gauge group 1:148
 - gauge invariance 1:145
 - gauge potential 1:145
 - gauge potential and field strength 1:147
 - Lie-algebra-valued p -forms 1:147
 - Lorentz force equation 1:142
 - Maxwell equations 1:144
 - nonabelian gauge theories 1:146
 - overview 1:141
 - relativistic kinematics 1:141
 - Yang–Mills equation 1:148
- and geometric measure theory 2:524
- Differential Galois theory, and Hamiltonian system
 - nonintegrability 2:629
- Differential geometry
 - classical differential geometry 3:78
 - cohomology 1:39
 - connections 1:37
 - curvature 1:38
 - curves and surfaces 1:33
 - fiber bundles 1:36
 - index theory 1:39
 - and integrable systems 3:78
 - metric properties 1:35
 - operators 1:37
 - overview 1:33–40
 - transformation groups 1:35
- Differential graded algebras, and deformation problem 2:21
- Differential operators, and Quillen determinant bundles 4:319
- Differential structures, structure on algebra, \ast -calculus 4:238
- Diffraction, and wave equations
 - geometrical optics 5:404
 - harmonic analysis 5:404
 - overview 5:401–407
 - propagation phenomena 5:402
- Diffusion
 - and Lagrangian dispersion 3:256
 - and multiscale approach 3:477
 - in phase space 1:24
- Diffusion operators, from scale spaces 3:6
- Diffusion speed, and Hamiltonian system instability 2:635
- Diffusivity, in turbulent flows 5:296
- Dilogarithm
 - and conformal field theory 3:601
 - in number theory 3:601
- Dimension functions, and Plancherel measure 4:350
- Dimensionless local balances, in thermohydraulics 3:495
- Dimension reduction
 - anti-self-dual Yang–Mills equation 3:47
 - and gamma-convergence 2:453
- Dimension regularization
 - and minimal subtraction 4:402
 - and perturbation theory 4:33
- Dimensions
 - bifurcation theory examples 1:275
 - bounded memory, finitely correlated pure states 2:339
 - Clifford algebra representations 1:525
 - Falicov–Kimball model 2:286
 - finite covering, spaces, and vector bundles 1:491
 - integrable evolution equations 1:348
 - $N=1$ supersymmetry 1:593
 - reduction, in fluid dynamics 1:283
 - and singular homology 1:547
 - small, isomorphisms, classical groups 1:503, 1:508
- Dimer covering, definition 2:61
- Dimer problems
 - combinatorics 2:62
 - Gibbs measures 2:61
 - overview 2:61–67
 - scaling limits 2:65
 - thermodynamic limit 2:63
- Diode model, random walks in random environments 4:362
- Dirac algorithm, and constrained systems 1:612
- Dirac brackets, and second-class constraints 1:614
- Dirac equation
 - in curved spacetime 2:69
 - free one-particle equation 2:75
- Dirac fermions
 - class AIII 5:212
 - classes BDI and CII 5:212
 - overview 5:211
- Dirac fields
 - free field properties 2:78
 - overview 2:67–74, 2:74–86, 2:83
- Dirac Hamiltonian, and fermions 1:323
- Dirac operator
 - and index theory 2:81
 - overview 2:74–86
- Dirac q -monopole, and quantum group bundles 4:242
- Dirac representations, Clifford algebras 1:525
- Dirac theory
 - constrained systems 1:412
 - one-particle theory 2:80
- Directional detection, wavelets applications 5:424
- Direct numerical simulations, for turbulence theories 5:298
- Direct sum
 - in quantum relative entropy 4:179
 - vector bundles 1:490
- Dirichlet boundary condition, definition 1:382
- Dirichlet branes
 - and AdS–CFT 1:175
 - and black hole statistical mechanics 1:377
 - coincident D3 branes 1:175
 - fractional D3 branes 2:464
 - and homological mirror symmetry 2:384
 - intersecting brane worlds 5:107
 - near-extremal D3 branes 1:176
 - and noncommutative soliton 3:522
 - and operads 3:615
 - phenomenology 5:106
 - and superstring theories 5:136
 - type IIB orientifolds 5:107

- Dirichlet branes (*continued*)
 and vacuum string field theory 5:102
 wrapped D5 branes 2:466
- Dirichlet problem, for Poisson
 equation 2:221
- Discrete dynamical systems, overview 3:59
- Discrete geometry, and integrable systems
 discrete surfaces 3:84
 overview 3:78–87
 quadrilateral lattice 3:79
- Discrete integrable surfaces
 asymptotic lattices and reductions 3:85
 isothermic nets 3:84
 Koenigs lattice 3:85
 2D Schrödinger equation 3:86
- Discrete isothermic nets,
 characteristics 3:84
- Discretely self-similarity, and gravitational
 collapse 1:670
- Discrete nonlinear Schrödinger systems,
 characteristics 3:559
- Discrete semigroups, renormalization
 group as 2:273
 critical field theory 2:276
 discrete semigroup 2:273
 eigenfunctions 2:274
 Gaussian measure perturbation 2:275
 properties 2:274
 ultraviolet cutoff removal 2:276
 unstable fixed points 2:277
- Discrete symmetries, characteristics 5:169
- Discrete systems
 and delta-bar problem method 2:38
 integrable systems 3:59–65
- Discrete translation-invariant
 determinantal random point fields,
 characteristics 2:50
- Discrete two-dimensional Schrödinger
 equation, and quadrilateral
 lattice 3:86
- Discrete wavelets
 fast wavelet transform 5:431
 frames 5:429
 higher dimensions 5:432
 and multiresolution
 analysis 5:429
 selection 5:431
- Discretization
 infinite Toda lattice 5:238
 via Bäcklund transformations 1:246
 and wavelet computation 5:417
- Disordered superconductors
 class C, 5:211
 class CI 5:211
 class D, 5:210
 class DIII 5:210
 Nambu space 5:209
 overview 5:209
- Disordered systems, and random matrix
 theory 4:345
- Dispersion relations
 in classical domain 2:89
 and damped harmonic oscillator 2:90
 energy–momentum space analyticity
 domains 2:97
 examples 2:88
 forward scattering case 2:96
 Kramers–Krönig relation 2:91
 Lagrangian dispersion 3:255–260
 linear dispersion 4:14
 overview 2:87–101
 QFT derivation 2:94
 and relativistic QFT 2:91
 relativistic quantum fields 2:95
 typical relations 2:88
- Dispersive property, kinetic
 equations 3:207
- Displacement current, electromagnetism
 overview 1:50
- Dissipative dynamical systems, infinite
 dimensions
 exponential attractors 2:107
 global attractor dimension 2:105
 global attractor generalization 2:108
 global attractors 2:104
 global Lyapunov functions 2:106
 inertial manifolds 2:107
 Kolmogorov’s entropy 2:108
 overview 2:101–110
 spatial dynamics and chaos 2:109
 spatio-temporal chaos 2:109
- Dissipative fluids, Einstein
 equations 2:198
- Dissipative operators, in Hilbert
 spaces 2:266–267
- Dissipative water-wave models,
 overview 2:136
- Distillable entanglement
 definition 2:234
 properties 2:235–236
 and quantum capacity 1:427
- Distinguished integrable reductions,
 multidimensional quadrilateral
 lattices 3:83
- Distributional drift, and stochastic
 differential equations 5:69
- Distribution function
 classical gravitational many-body
 problems 2:579
 kinetic theory, Einstein
 equations 2:199
- Divergences
 classical mechanics overview 1:18
 magnetic field 1:44
- Division algebras, vector fields 1:528
- Divisors, and Riemann surfaces 4:422
- DLR, and phase transitions 4:53
- DNS *see* Direct numerical simulations
- DOC *see* Domain of outer
 communications
- DOD *see* Difference of two Diracs
- Dolbeault complex
 characteristics 3:25
 index 3:28
- Domain of outer communications, and
 stationary black holes 5:42
- Domains, in C^n
 analytic continuation
 theorems 4:541
 bounded domains and
 automorphisms 4:542
 Cartan–Thullen theorem 4:542
 Levi problem 4:542
 1D theory extension 4:540
- Dominated splittings
 C^1 -generic diffeomorphisms 2:501
 and partial hyperbolicity 2:726
- Donaldson invariants, and
 instantons 2:457
- Donaldson–Witten theory
 Donaldson’s version 2:111
 extensions 2:116
 overview 2:110–117, 5:282
 results 2:113
 u -plane integral
 applications 2:116
 definition 2:113
 Seiberg–Witten contribution 2:115
 wall-crossing formula 2:115
 various perspectives 5:287
 Witten’s version 2:112
- Double complex, spectral sequences
 4:632
- Doubling bifurcations, universal
 geometry 5:344
- Dressing method, and integrable
 systems 4:430
- Drinfeld associator, Kontsevich integral
 construction 3:237
- Drinfeld–Sokolov method, and
 solitons 4:599
- Drinfeld–Sokolov reduction, and
 bi-Hamiltonian formulation 1:295–296
- Dromions
 characteristics 3:97
 and integrable PDEs 3:116
- Droplets
 condensation, moderate
 deviations 3:262
 and nucleation 4:51
- Droplet–scaling picture, short-range spin
 glasses 4:572
- Drops, in capillary surfaces 1:443
- DS *see* Dynamical systems
- DSS *see* Discretely self-similarity
- dS spacetimes *see* De Sitter spacetimes
- Dual chain complexes, and
 cohomology 1:548
- Dual Hahn polynomials,
 characteristics 3:644

- Duality
 action-angle transforms in CMS systems 1:409
 and algebraically completely integrable 3:77
 and branes 1:375
 in convex analysis 1:644
 convex variational problems 1:647
 definition 1:382
 eigenfunction transforms in CMS systems 1:410
 functional analysis overview 1:90
 large- N duality 3:266
 in mass transport problems 1:648
 and M theory 5:137
 and percolation in two dimensions 4:24
 Poisson–Lie T -duality 1:273
 and quantum groups 2:687
 Seiberg–Witten duality 4:507
 self-duality 4:114
 state-channel duality 1:427
 in topological QFT, overview 2:118–125
 and Wu–Yang criterion 2:204
- Duistermaat–Heckman formulas, and equivariant cohomology 2:249
- Dupré apparatus, and capillary surfaces 1:436
- Dust, Einstein equations 2:196
- Dyck path, in combinatorics 1:555
- Dynamical Boltzmann–Einstein formula, and fluctuations 3:359
- Dynamical localization, for quasiperiodic potentials 3:337
- Dynamical map, and quantum field equations 4:226
- Dynamical rearrangement, symmetry 4:223, 4:227
- Dynamical semigroups *see* Quantum dynamical semigroups
- Dynamical systems
 in black hole mechanics 1:304
 C^r -generic systems 2:498
 discrete 3:59
 dissipative dynamical systems 2:101–110
 dissipative water-wave models 2:136
 and equilibrium states 3:418
 fractal dimensions
 box dimensions 2:395
 embedology 2:400
 future work 2:401
 gauge functions 2:395
 geometric constructions 2:396
 Hausdorff dimension 2:395
 Hénon attractor 2:399
 hyperbolic measures 2:397
 infinite-dimensional dynamical systems 2:400
 Julia and Mandelbrot sets 2:400
 local Lyapunov dimension 2:398
 logistic map 2:398
 Lorenz attractor 2:399
 multifractal analysis 2:397
 overview 2:394–402
 probability 2:401
 Smale horseshoe 2:398
 spiral trajectories 2:400
 thermodynamic formalism 2:396
 generic properties overview 2:494–502
 C^1 -generic systems 2:499
 glassy disordered systems 2:555
 and gravitational collapse 1:669
 Hamiltonian systems in
 hydrodynamics 2:135
 and Hamilton–Jacobi equations, overview 2:636–644
 hydrodynamics asymptotical models 2:133
 hyperbolic systems 2:721–729
 infinite-particle systems 3:540
 integrable systems 3:107
 liquid crystals 3:324
 long-time behavior 3:541
 loop quantum gravity 3:342
 mean-field limit 3:542
 nonequilibrium models 3:543
 peakon, CH equation 4:15
 perturbative gauge theory 3:521
 QFT 4:221
 quantum mechanics overview 1:116
 random maps 4:331
 random noise 4:330
 random noise framework 4:332
 and separatrix splitting 4:539
 stochastic water-wave models 2:137
 symmetry and symmetry breaking
 bifurcations 5:189
 equivariant bifurcation theory 5:187
 equivariant dynamics 5:185
 Hopf bifurcation and mode interactions 5:189
 isotropy subgroups and commuting linear maps 5:184
 overview 5:184–190
 and symplectic reduction 5:193
 and thermodynamics
 Boltzmann’s and Nernst’s work 2:131
 FPU phenomenon 2:127
 nearly integrable *vs.* hyperbolic systems 2:126
 overview 2:125–133
 water waves example 2:133–138
 water waves initial-value problem 2:134
- Dynamical zeta functions, regularization
 Grothendieck–Fredholm case 4:387
 and intermittency 4:388
 and kneading determinants 4:389
 overview 4:386–391
- Dynamics concepts
 asymptotic dynamics 5:234
 classical dynamics 4:513
 computational fluid dynamics 5:298
 contour dynamics 5:397
 cosmological dynamics 1:371
 critical dynamics 3:134
 equilibrium dynamics 2:557
 equivariant dynamics 5:185
 fluid dynamics 1:281–285
 geophysical dynamics 2:534–539
 glassy dynamics 3:132
 Glauber dynamics 3:131
 Hamiltonian dynamics 4:301
 Hamiltonian fluid dynamics 2:593–600
 holomorphic dynamics 2:652–659
 hydrodynamics 1:309, 2:133, 3:362, 3:376
 interface dynamics 3:132
 Kawasaki dynamics 3:131–132
 Kawasaki–Glauber dynamics 3:365
 London quantum hydrodynamics 5:117
 low temperature dynamics 3:133–134
 macroscopic dynamics 3:358
 magnetohydrodynamics 3:375–383, 3:375
 microscopic dynamics 3:358
 Newtonian dynamics 1:41
 off-equilibrium dynamics 2:557
 phase transition dynamics 4:47–52
 point-vortex dynamics 4:66–79
 quantum dynamics 2:435, 4:165–168
 topological dynamics 5:343
 topological vortex dynamics 5:271
 vortex dynamics 5:390–399
- Dynkin diagrams, overview 3:296
- Dyson’s threefold way
 class A, 5:207
 classes AI and AII 5:208
 overview 5:207
- ## E
- EA model *see* Edwards–Anderson Ising spin glass model
- ECT *see* Einstein–Cartan theory
- Edge detection, wavelets
 applications 5:423
- Edge scaling, Plancherel measure 4:348
- Edge states, and fractional quantum Hall effect 2:410
- Edwards–Anderson Ising spin glass model
 low-temperature structure 4:572
 open questions 4:574
- EDYM equations *see* Einstein–Dirac–Yang–Mills equations
- Effective actions, and string theory
 phenomenology 5:109
- Effective diffusion
 for passively advected scalar 3:481
 in periodic potential 3:480
 in porous medium 3:480

- Effective field theory
 chiral perturbation theory 2:145
 classification 2:139
 compactification, overview 1:586
 electroweak symmetry breaking 2:141
 heavy-quark effective theory 2:142
 and noncommutative spacetime 2:141
 nonrelativistic quantum
 chromodynamics 2:144
 and nuclear physics 2:146
 overview 2:139–147
 soft-collinear effective theory 2:143
 standard model as 2:140
 standard model at low energies 2:144
- Effective magnetic field, and fractional quantum Hall effect 2:409
- Effective potentials, in constructive QFT 1:625
- Effective quantity, and multiscale approach 3:476
- EFQT *see* Euclidean quantum field theory
- EFT *see* Effective field theory
- Egorov's theorem, h -pseudodifferential operator application 2:709
- Eigenfunctions
 in arithmetic quantum chaos 1:218
 in Calogero–Moser–Sutherland systems 1:410
 ergodic, problems and results 4:191
 and h -pseudodifferential operators 2:703
 quantum chaotic systems 4:194
 quantum completely integrable systems
 Birkhoff normal forms 2:150
 blow-up 2:149
 lower bounds 2:151
 modes and quasimodes 2:152
 open problems 2:154
 overview 2:148–155
 QCI background 2:149
 upper bounds 2:154
 and quantum ergodicity 4:185
 and renormalization groups 2:274
 and Schrödinger operators 4:488
- Eigenvalues
 clustering, and h -pseudodifferential operators 2:710
 and quantum ergodicity 4:185
 semi-classical trace formula 4:517
 separation problems 4:533
- Eight-vertex model
 commuting transfer matrices 2:158
 corner transfer matrices 2:162
 definition 2:156
 one-point functions 2:162
 overview 2:155–164
 parameterization 2:159
 Yang–Baxter equations 2:158
- Eilenberg–Moore spectral sequence, properties 4:632
- Einstein–Cartan theory
 and gravitation
 Bianchi identities 2:193
 generalized Mathisson–Papapetrou equation of motion 2:193
 and general relativity theory 2:194
 identity from local invariance 2:192
 projective transformations 2:192
 Sciama–Kibble field equations 2:193
 spinning fluid 2:193
 spin and torsion 2:194
 history 2:189
 Hodge duals 2:191
 linear connection 2:191
 metric-affine geometry 2:192
 notation 2:189
 overview 2:189–195
 physical motivation 2:190
 Riemann–Cartan geometry 2:192
 tensor-valued differential forms 2:190
- Einstein–Dirac–Yang–Mills equations
 black-hole solution attempts 2:70
 particle-like solutions 2:71
 spherically symmetric, overview 2:69
- Einstein equations
 black-hole solutions 2:68
 charged fluids 2:199
 and Dirac fields 2:67
 dissipative fluids 2:198
 dust 2:196
 elastic media 2:200
 exact solutions
 algebraic classification 2:165
 anti-de Sitter spacetimes 2:166
 cosmological models 2:172
 de Sitter spacetimes 2:166
 and material sources 2:171
 Minkowski spacetimes 2:166
 overview 2:165–173
 Reissner–Nordström metrics 2:167
 Schwarzschild metrics 2:167
 stationary axisymmetric solutions 2:169
 symmetry classifications 2:166
 fluid sources overview 2:196
 Friedman–Robertson–Walker metric 4:562
 and general relativity 2:67, 2:492
- Hamiltonian reduction
 Bianchi models 2:618
 cosmological applications 2:622
 field equations 2:611
 overview 2:607–623
 reduced Hamiltonian 2:614
 topological background 2:607
- initial-value formulation, conformal method
 closed manifold CMC data 2:177
 maximal asymptotically Euclidean data 2:178
 overview 2:176
- initial-value formulation, constraint equations
 conformal thin sandwich 2:178
 gluing solutions 2:180
 parabolic methods 2:179
 quasispherical ansatz 2:179
 space + time field
 decomposition 2:174
- initial-value formulation, overview 2:173–182
- and knot invariants 3:218
- mathematical cosmology 1:653
- with matter, kinetic models
 Boltzmann equation 2:199
 distribution function and moments 2:199
 Liouville–Vlasov equation 2:199
- with matter, overview 2:195–200
- and Minkowski space stability 5:15
- perfect fluid 2:196
- radiative solutions
 boost-rotation symmetric spacetimes 2:171
 cylindrical waves 2:170
 plane waves 2:170
 Robinson–Trautman solutions 2:171
- and symmetric-hyperbolic systems 5:165
- in 3 + 1 form
 boundary conditions 1:610
 Cauchy problem 1:609
 constraint equations 1:608
 evolution equations 1:609
- Einstein equivalence principle
 local Lorentz invariance tests 2:483
 local position invariance tests 2:483
 overview 2:482
 tests 2:482
- Einstein gravity, canonical structure
 classical constraints 5:455
 constraint algebra 5:456
 overview 5:454
- Einstein–Hilbert action
 definition 1:382
 overview 5:454–455
- Einstein manifolds
 analytical methods 2:188
 basic properties 2:183
 and conformal field theory 1:180
 Einstein–Sasaki structures 2:186
 exceptional holonomy 2:186
 higher cohomogeneity 2:187
 homogeneous equations 2:187
 hyper-Kähler manifolds 2:185
 Kähler manifolds 2:184
 obstructions 2:183
 overview 2:182–189
 quaternionic Kähler manifolds 2:186
 3-Sasakian structures 2:186

- special holonomy 2:184
- uniqueness and moduli 2:184
- Einstein–Podolski–Rosen–Bohm theorem, and quantum mechanics foundation 4:262
- Einstein–Podolski–Rosen paradox, and entanglement history 2:228
- Einstein radius, in gravitational lensing 2:569
- Einstein rings, in gravitational lensing 2:574
- Einstein–Sasaki structures, description 2:186
- Elasticity, liquid crystal phases 3:322
- Elastic media, Einstein equations 2:200
- Electric–magnetic duality gauge invariance 2:202
- monopoles 2:202
- nonabelian duality treatments 2:208
- nonabelian duality via loop variables 2:205
- overview 2:201–208
- sources 2:202
- and Wu–Yang criterion 2:204
- Electric networks, and Yang–Baxter equations 5:465
- Electrodynamics *see* Quantum electrodynamics
- Electromagnetic fields *see also* Magnetic fields in classical physics 1:191
- and Minkowski spacetime 1:105
- and quantum electrodynamics 4:168
- in quantum physics 1:193
- Electromagnetism charge conservation 1:47
- charges, fields, Lorentz-force law 1:42
- continuity equation 1:47
- displacement current 1:50
- Faraday’s law of induction 1:45
- field of charges in uniform motion 1:46
- Galilean invariance limits 1:45
- Galilean relativity inconsistency 1:44
- gauge transformations 1:50
- and gravitational waves 2:582
- historical overview 1:50
- magnetic field divergence 1:44
- in magnetohydrodynamics description 3:376
- Maxwell’s equations 1:46
- monochromatic plane waves 1:48
- overview 1:40–51
- potentials 1:49
- relativity in Newtonian dynamics 1:41
- source-free equations 1:48
- stationary charge distributions 1:43
- and Yang–Mills equation 1:149
- and Yang–Mills theories 5:167
- Electron motion, in magnetic field 4:246
- Electroweak interactions, in standard model 5:34
- Electroweak symmetry breaking, and standard model 2:141
- Electroweak theory *see also* Glashow–Weinberg–Salam theory extension 2:214
- overview 2:209–216
- Elementary moves, finite-type invariants 2:349
- Elementary quantum mechanics Malliavin calculus applications 3:387
- symmetries 5:179
- Elementary random algebraic geometry, overview 4:323
- Elitzur’s theorem, and symmetry breaking 5:202
- Elliptic cohomology, overview 1:552
- Elliptic complexes, classical 3:24
- Elliptic differential equations, linear theory biharmonic problem 2:222
- Dirichlet problem for Poisson equation 2:221
- general elliptic problems 2:223
- Lax–Milgram theorem 2:220
- linearized elasticity 2:227
- method 2:217
- mixed boundary conditions 2:222
- motivation problem 2:216
- Neumann problem for Poisson equation 2:221
- Newton problem for general operators 2:222
- overview 2:216–228
- second-order elliptic problems 2:223
- Sobolev spaces of positive order 2:217
- Stokes system 2:226
- unbounded open 2:225
- variational approach 2:220
- Elliptic equations, for fluid mechanics 2:366
- Elliptic fibrations, relative Fourier–Mukai transforms 2:380
- Elliptic–hyperbolic motion, classical gravitational many-body problems 2:578
- Elliptic hypergeometric series Bailey’s $_{10}W_9$ transformation formula 4:115
- Jackson’s ${}_8W_7$ formula 4:115
- overview 4:115
- Ruijsenaars’ gamma function 4:115
- Elliptic theory, and h -pseudodifferential calculus 2:705
- ELs *see* Evolution laws
- Embedded minimal disks, and minimal submanifolds 3:425
- Embeddings bounding tori 1:487
- in space BV 2:412–413
- spacetime boundaries from classifications 1:328
- conformal embeddings 1:328
- Minkowski space 1:326
- overview 1:326
- Embedology, in dynamics 2:400
- EMFs *see* Electromagnetic fields
- Energy conservation, and general relativity 2:493
- Energy-equipartition thresholds, and Fermi–Pasta–Ulam phenomena 3:547
- Energy estimation in fluid mechanics 2:365
- magnetic knots and links 5:276
- and viscous incompressible fluids 5:371
- Energy functional, for Ginzburg–Landau model 5:355
- Energy identity, in symmetric–hyperbolic systems 5:162
- Energy method and capillary surfaces 1:433
- in turbulence 5:351
- Energy–momentum space analyticity domains 2:97
- complex, analyticity in 2:95
- QFT analytic program linear program results 4:471
- nonlinear program 4:473
- N -point structure functions 4:468
- off-shell character 4:470
- scattering kernels 4:469
- Energy–momentum tensor, and Einstein–Cartan theory 2:194
- Energy relaxation, in nonlinear chains 3:546
- Energy–time coordinates, classical mechanics overview 1:6
- Ensembles equilibrium statistical mechanics overview 1:54
- equivalence 1:56
- Entanglement collection 2:235
- comparative entanglement 2:233
- distillable, and quantum capacity 1:427
- Gaussians 2:232
- historical development 2:228
- multipartite stars 2:232
- open problems 2:232
- orthogonally invariant states 2:231
- overview 2:228–233, 2:233–236
- properties 2:235
- qualitative theory 2:229
- qubits 2:231
- Entanglement-assisted classical capacity, overview 1:418
- Entanglement cost definition 2:234
- properties 2:235–236

- Entanglement-enhanced quantum capacity
 notation 1:418
 overview 1:418–423
 protocol relationships 1:422
 and quantum reverse Shannon
 theorem 1:421
- Entanglement of formation,
 properties 2:235–236
- Entropy
 and arithmetic quantum chaos 1:217
 in basic convex analysis 1:644
 and D-branes and black branes 1:377
 entanglement 2:235–236
 equilibrium statistical mechanics
 overview 1:53
 and ergodic theory 2:254
 and Hamilton–Jacobi equation 3:363
 historical overview 4:177
 and holomorphic dynamics 2:659
 and homoclinic points 2:676
 and interacting particle systems 3:124
 and inviscid compressible fluid
 flow 1:601
 and kinetic equations 3:207
 kinetic equations 3:201
 macroscopic, definition 1:382
 and nonequilibrium statistical
 mechanics 3:534
 and quantitative transversality
 almost-transversality 2:240
 characteristics 2:238
 definitions 2:238
 examples 2:240
 as generic situation 2:239
 measures and dimensions 2:239
 overview 2:237–242
 quantum dynamical semigroups 4:163
 and quantum sources 4:611
 random dynamical systems 4:335
 and source coding 4:609
 and symmetrization 5:161
- Enumerative geometry, as intersection
 theory application 3:154
- Enzymatic catalysis, and multiscale
 approaches 3:475
- EPR–Bohm theorem *see*
 Einstein–Podolski–Rosen–Bohm
 theorem
- EPR paradox *see*
 Einstein–Podolsky–Rosen paradox
- Equations of motion
 and fluid flow stability 5:1
 Lagrange and Hamilton forms 1:3
 nonabelian BF theories 1:258
 in turbulence theories 5:297
- Equations of state, perfect fluid 2:197
- Equilibrium crystals, statistical
 mechanics 5:61
- Equilibrium dynamics, and glassy
 disordered systems 2:557
- Equilibrium states
 and dynamics 3:418
 and equivariant dynamics 5:185
 and ergodic theory 2:253
 and interacting stochastic particle
 systems 3:133
- Equilibrium statistical mechanics
 atomic dimensions 1:62
 atoms and molecules 1:51
 coalescence catastrophe 1:84
 continuous symmetries 1:65
 critical points 1:76
 droplet condensation deviations 3:262
 ensembles 1:54
 equivalence of ensembles 1:56
 evaporation catastrophe 1:85
 ferromagnetic Ising model 1:72
 finite-volume effects 1:71
 fluctuations 1:77
 heat and entropy 1:53
 heat theorem and ergodic
 hypothesis 1:53
 infrared inequality 1:86
 large deviations 3:261–263
 lattice models 1:68
 overview 1:51–88
 phase coexistence geometry 1:74
 phase separation and coexistence 1:75
 phase separation line and surface
 tension 1:75
 and phase transition absence 1:64
 phase transitions and boundary
 conditions 1:59
 pressure, temperature, kinetic
 energy 1:52
 quantum statistics 1:82
 renormalization group 1:79
 subadditivity method 1:85
 symmetry-breaking phase
 transitions 1:70
 temperature and density 1:67
 thermodynamic functions 1:58
 thermodynamic limit 1:57
 thermodynamic limits and
 inequalities 1:69
 van der Waals theory 1:63
 virial theorem 1:62
- Equivalence
 n -equivalence 2:345
 and general relativity 2:487
 nonequilibrium ensembles 3:537
 in optical caustics 3:625
 transformations in Fedosov
 quantization 2:295
- Equivariant bifurcation theory
 asymptotic stability 5:188
 axial isotropy subgroups 5:187
 branching patterns and finite
 determinancy 5:188
 and dynamics 5:188
- nonaxial maximal isotropy
 subgroups 5:187
 overview 5:187
 submaximal isotropy subgroups 5:188
- Equivariant cohomology
 Borel’s model 2:242
 and Cartan model, overview 2:242–250
 characteristic forms 2:246
 de Rham theory 2:245
 differential algebras 2:243, 2:245
 and Duistermaat–Heckman
 formulas 2:249
 index theory 2:249
 localization theorem 2:248
 Mathai–Quillen formalism 3:392, 3:395
 Thom forms 2:247
 Weil algebra 2:244
- Equivariant dynamics
 equilibria 5:185
 and forced symmetry breaking 5:187
 periodic solutions 5:185
 relative equilibria and skew
 products 5:185
 relative periodic solutions 5:186
 robust heteroclinic cycles 5:186
 and symmetric attractors 5:186
- Equivariant Hopf bifurcation, and mode
 interactions 5:189
- Equivariant problems, classical Lefschetz
 formula 3:31
- Equivariant vector bundles,
 Mathai–Quillen formalism 3:393
- Ergodicity
 decomposition and finitely correlated
 states 2:337
 and determinantal random fields 2:51
 entropy and reversibility 2:254
 equilibrium states and mixing
 properties 2:253
 equilibrium statistical mechanics
 overview 1:53
 for harmonic functions on Young
 graph 4:352
 induced transformations 2:252
 multiplicative ergodic theorem 3:350
 nonergodicity 4:377
 overview 2:250–256
 quantum dynamical semigroups 4:162
 quantum ergodicity 4:183–196
 quantum unique ergodicity 1:216
 and random dynamical
 systems 4:332
 in random matrix theory 4:341
 symbolic coding 2:725
 theorems 2:251
 tower constructions 2:252
- Ergodic quantum sources, and data
 compression 4:614
- Ernst equation, Einstein equation
 solutions 2:169

- Error correction, quantum *see* Quantum error correction
- Essential boundary conditions, in fluid mechanics 2:365
- Euclidean data, maximal asymptotically, Einstein equations 2:178
- Euclidean quantum field theory
interacting fields 2:261
model definition 4:482
one-particular systems 2:257
overview 1:617, 2:256–265
physical interpretation 2:264
second quantization and free fields 2:260
- Euclidean space, instantons on 3:45
- Euclidean stochastic equations, and QFT with indefinite metric 4:217
- Euclidean-type Sobolev inequalities, properties 3:35
- Euler classes
characteristic classes 1:495
index theorems 3:26
in Mathai–Quillen construction 3:391
- Euler equations
compressible inviscid fluids 3:165
and fluid flow stability 5:5
inviscid compressible fluid flows
adiabatic equations 1:599
entropy solutions 1:601
isentropic equations 1:598
local well-posedness 1:597
multidimensional steady theory 1:599
multidimensional unsteady problems 1:600
overview 1:596
singularity formation 1:597
local existence and blow-up problem 3:12
overview 3:10–17
as partial differential equations 4:7
perfect fluid 2:196
stochastic hydrodynamics 5:73
2D, weak solutions 3:14
- Eulerian action principles, in Hamiltonian fluid dynamics 2:597
- Eulerian variables, and Hamiltonian fluid state 2:595
- Euler–Poincaré–Hamel action, in Hamiltonian fluid dynamics 2:598
- Euler–Poincaré theory, peakon solutions of CH equation 4:16
- Euler’s integral representation, q -analogue 4:107
- Euler top, as Hamiltonian system example 3:460–461
- Evaporation catastrophe, equilibrium statistical mechanics overview 1:85
- Even dimensions, Clifford algebra representations 1:525, 1:526
- Even theory
overview 4:481
two-particle bound states 4:485
- Event horizons, in black hole mechanics 1:305
- Evolution domain, and multidimensional nuclear magnetic resonance 3:597
- Evolution equations
boundary-value problems 1:350
canonical general relativity 1:416
for fluid flow stability 5:3
and general relativity theory 1:609
linear type 2:265
nonlinear type 2:270
overview 2:265–272
and point-vortex dynamics 4:67
- Evolution laws
adiabatic and unitary evolution 2:530
and cellular automata 1:456, 1:457
Markovian evolution 3:358
and multidimensional cellular automata 1:459–460
quantum evolution 2:708
stochastic Loewner evolutions 5:80
- Evolution operators
and classical mechanics 2:707
for finitely correlated states 2:336
- Exact couples, spectral sequences 4:625
- Exact renormalization group, as continuous semigroup 2:277
- Exceptional holonomy groups
compact G_2 -manifolds 4:444
group G_2 , 4:443
group $\text{spin}(7)$ 4:444
overview 4:441–446
 $\text{spin}(7)$ -manifolds 4:444
- Excision
 C^* -algebras 3:250
in K -theory 3:247
and twisted K -theory 3:252–253
- Excitation gases, local response 5:119
- Excited states, quantum Calogero–Moser systems 4:126
- Existence in law, stochastic differential equations 5:67
- Exotic structures, on open 4-manifolds 2:392
- Expansions
asymptotic expansion 4:37
bulk–boundary operator product expansion 1:339
cluster expansions 1:629–630, 1:531–537
cyclotomic expansion 2:354
large- N expansion 3:263–268
matched expansions 3:465–466
Mayer expansion 1:531
operator product expansion in QFT 3:616–620
perturbative expansion 2:354, 3:618
semiclassical expansions 2:441
Weyl expansion 4:516
- Explicit Hamiltonians, for Hitchin systems 3:75
- Exponential attractors, dissipative dynamical systems 2:107
- q -Exponential series, equations 4:106
- Exponential series, example 3:639
- Exponential stability, and Hamiltonian systems 2:631
- Extended algebras, and boundary conformal field theory 1:339
- Extended supergravities, overview 5:126
- Extended supersymmetric
compactification, in superstring theory 1:590
- Extended supersymmetry
construction 5:149
twisting, and TQFT 5:281
- Exterior domain, unbounded open 2:225
- Exterior product, and geometric measure theory 2:523
- External fields
boson and fermion field algebras 1:322
boson and fermion overview 1:318–326
bosons 1:324
fermions 1:323
- Extra-dimensional volume, in brane worlds 1:368
- Extrudate swell, as non-Newtonian behavior 3:561
- ## F
- Factorization, and perturbative quantum chromodynamics 4:150
- Factorized Hamiltonian, quantum Calogero–Moser systems 4:125
- Faddeev–Popov trick, and abelian BF theories 1:260
- Falicov–Kimball model
canonical ensemble 2:284
definitions 2:283
grand-canonical ensemble 2:285
grand-canonical phase diagram 2:287
ground-state configuration dimension 2:288
model symmetries 2:286
overview 2:283–291
segregated state 2:286
- Family catastrophes, and optic caustics 3:623
- Faraday’s law of induction, electromagnetism overview 1:45
- Fast motions, decoupling in averaging methods 1:228
- Fast–slow systems, and bifurcation theory 4:593

- Fast variables, elimination in averaging
methods 1:228
- Fast wavelet transform, discrete
wavelets 5:431
- Fatou's conjecture, and hyperbolic
maps 2:656
- Fatou sets, holomorphic dynamics 2:654
- Fault tolerance
and error correction 4:200
and quantum error-corrected
codes 4:199
threshold 4:200
- Fault-tolerant gates, and quantum
error-corrected codes 4:199
- FCDs *see* Focal conic domains
- Fedosov quantization
adaptations 2:297
automorphisms 2:295
derivations 2:295
equivalence transformations 2:295
generalizations 2:297
modifications 2:297
overview 2:291–300
star products 2:292
- Feigenbaum attractor, universal
geometry 5:344
- Feigenbaum–Coullet–Tresser
renormalization conjecture, and
universality phenomenon 5:345
- Feigenbaum Julia sets, and
renormalization 5:347
- Fermat's theorem, in gravitational
lensing 2:571
- Fermi gas model, physical
observables 2:301
- Fermi interactions, and
Glashow–Weinberg–Salam
theory 2:211
- Fermion Chern–Simons theory,
overview 2:407
- Fermion doubling, in lattice gauge
theories 1:211
- Fermion fields, and perturbation
theory 4:33
- Fermion gas, and determinantal random
point fields 2:49
- Fermionic Brownian motion, and path
integrals 4:11
- Fermionic systems
in external potential 2:302
free Fermi gas 2:301
interacting systems 2:303
Landau Fermi liquids 2:304
1D systems 2:305
overview 2:300–307
quantum statistics 2:300
second-quantization formalism 2:301
superconductors 2:306
- Fermions
boson–fermion equivalence 5:331
bosonization–fermionization
model 5:328
degeneracy 5:116
Dirac fermions 5:211
in external fields, overview 1:318–326
in external potential 2:302
field theory 2:281
in four dimensions 3:572
and lattice gauge theory 3:277
many-body formalism 1:323
nonperturbative renormalization 4:413
one-particle formalism 1:323
perturbative renormalization 4:411
and random matrix theory
supersymmetry method 5:153
relativistic, field algebras and quasifree
representations 1:322
renormalization overview 4:411
in three dimensions 3:577
in two dimensions 3:576
- Fermion theory, composite-fermion
theory 2:406
- Fermi–Pasta–Ulam phenomena
and Boltzmann's and Nernst's
work 2:131
conceptual development 2:127
energy-equipartition thresholds 3:547
and heat transport 3:549
numerical experiment 3:545
resonance-overlap criterion 3:546
- Fermi sea, at half filling 2:409
- Fermi surface
quantum statistics overview 1:82
renormalization 4:411
- Ferromagnetic Ising model, equilibrium
statistical mechanics overview 1:72
- Ferromagnetism
and Hubbard model 2:715
superconductivity unification via $SO(5)$
theory 2:648
- Feynman diagrams
classical mechanics overview 1:17
in number theory 3:602
spin foams as 4:654
- Feynman path integrals
analytic continuation 2:309
Fourier transform approach 2:309
infinite-dimensional oscillatory
integrals 2:310
Malliavin calculus applications
3:388
on manifold 4:9
mathematical definitions 2:308
overview 2:307–313, 4:516
semiclassical limit 4:516
solution approaches 2:312
trace and loop manifolds 4:516
white-noise calculus 2:312
- Feynman rules, in perturbation
theory 4:30
- Fiber bundles, differential geometry
overview 1:36
- Fibonacci numbers, and
combinatorics 1:555
- Fibonacci tiling, fundamental
domain 4:314
- Fibrations, elliptic, relative Fourier–Mukai
transforms 2:380
- Field algebras, relativistic bosons and
fermions 1:322
- Field of charges, in uniform motion 1:46
- Field equations, and Hamiltonian
reduction 2:611
- Field strength, and gauge potential 1:147
- Field theories
and Batalin–Vilkovisky formalism 1:251
class field theory 3:603
and classical integrable systems 3:50
cohomological field theory 3:397
effective field theory 1:586, 2:139–147
Euclidean field theory 2:256–265
fermions 2:281
free field theory 3:617
infinite-dimensional classical field
theory 3:37
massive field theory 4:476
renormalization group 4:410
symmetry breaking
Elitzur's theorem 5:202
global symmetries 5:198
Goldstone bosons 5:200
overview 5:198–204
phase transitions 5:199
phase transitions and crossovers
5:203
spontaneously broken gauge
theories 5:201
- Filament methods, in vortex
dynamics 5:398
- Filling factor, and Landau levels 2:405
- Filtered chain complexes, spectral
sequences 4:625
- Filtered complexes, knot
homologies 3:209
- Filtered objects, spectral sequences 4:624
- Filtrations, and finite-type
invariants 2:349, 2:350
- Finite box, Ising model 3:261
- Finite cylinder, adiabatic piston 1:166
- Finite determinacy, and equivariant
bifurcation theory 5:188
- Finite-dimensional algebras
associative algebras 2:314
compact groups 1:584
overview 2:313–322
path algebras 2:314
related topics 2:321
Ringel–Hall algebras 2:317
- Finite-dimensional classical mechanics, and
Poisson manifolds 3:37

- Finite-dimensional Lie groups, infinite group comparison 3:40
- Finite-dimensional systems, and fluid Hamiltonian 2:598
- Finite dimensions, Quillen determinants in 4:315
- Finite-element approximation space, in fluid mechanics 2:368
- Finite extinction, for Ricci flow 3:431
- Finite groups, notions and group actions 2:322
- Finite group symmetry breaking
 - extensions and physical applications 2:327
 - finite group notions 2:322
 - G-invariant polynomials 2:324
 - group action geometry 2:324
 - Landau polynomial 2:325
 - Landau polynomial critical points 2:325
 - Landau potential reduction 2:326
 - nonstationary and nonvariational problems 2:326
 - orbit space geometry 2:325
 - overview 2:322–327
 - spontaneous breaking 2:323
- Finitely correlated states
 - correlation generation 2:334
 - ergodic decomposition 2:337
 - evolution operator construction 2:336
 - ground states and gaps 2:338
 - local support, asymptotic form 2:338
 - matrix product states 2:335
 - notation 2:334
 - overview 2:334–340
 - pure state density 2:339
 - pure states 2:337
 - pure state with bounded memory dimension 2:339
 - VBS construction 2:335
- Finite mass, in adiabatic piston
 - finite cylinder 1:168
 - first order equation 1:167
 - infinite cylinder 1:167
- Finite open-end Toda lattice
 - factorization solution 5:239
 - model 5:238
 - Moser's solution 5:239
- Finite periodic Toda lattice
 - inverse-spectral transformation 5:240
 - loop algebras 5:240
 - model 5:239
- Finite-type invariants
 - associators and trivalent graphs 2:344
 - bounds and computational results 2:342
 - chord diagrams 2:341
 - claspers 2:345
 - classical knot polynomials 2:341
 - commutators 2:345
 - configuration space integrals 2:344
 - n -equivalence 2:345
 - Gauss sums 2:346
 - independent modifications 2:345
 - Jacobi diagrams 2:342
 - Kontsevich integral 2:343
 - Kontsevich integral computation 2:346
 - Kontsevich integral taming 2:346
 - Melvin–Morton conjecture 2:347
 - non-knot concepts 2:347
 - overview 2:340–348, 2:341
 - perturbative Chern–Simons–Witten theory 2:344
 - polynomiality 2:346
 - proof 2:343
 - Rozansky–Witten theory 2:346
 - step-by-step integration 2:345
 - structure of \mathcal{A} 2:342
 - 3-manifolds
 - conventions and notations 2:348
 - FTI generality 2:349
 - FTIs for ZHS 2:349
 - Goussarov–Habiro definition 2:349
 - Habiro ring 2:354
 - Le–Murakami–Ohtsuki invariant 2:351
 - mathematical theories 2:348
 - overview 2:348–356
 - quantum invariants and perturbative expansion 2:352
 - Vassiliev perspective 2:345
 - volume conjecture 2:347
- Finite volumes
 - and equilibrium statistical mechanics 1:71
 - in fluid mechanics
 - Poisson problem 2:371
 - principles 2:370
 - Stokes equations 2:372
 - transport equation 2:371
- Finite Weyl systems
 - Clifford channels 2:330
 - overview 2:328–334
 - representations 2:329
 - stabilizer formalism 2:331
- First-order deformation, algebras 2:18
- First-order Lagrangians, in constrained systems 1:611–612
- First-passage percolation, characteristics 4:26
- Five-body problems, classical gravitational many-body problems 2:579
- Fixed-point theorems, for compact perturbations of identity 3:282
- Flag manifolds, overview 1:501
- Flat connections, and A–B geometry 1:196
- Flat surfaces, and polygonal billiards 4:86
- FL models *see* Friedmann–Lemaître models
- Floer homology
 - Arnol'd conjecture 2:358
 - in gauge theory 2:477
 - Heegaard Floer homology 2:364
 - instanton Floer homology 2:362
 - Morse theory 2:356
 - overview 2:356–365
 - Seiberg–Witten Floer homology 2:363
- Flows, random perturbations 4:331
- Fluctuation–dissipation theorem, generalized theorem 3:548
- Fluctuations
 - equilibrium statistical mechanics overview 1:77
 - and quantum central-limit theorems 4:136
 - thermodynamic fluctuations 3:535, 3:557–366
- Fluid dynamics
 - bifurcations in 1:282
 - and capillary surfaces 1:436
 - charged fluids 2:199
 - compressible fluids 5:165
 - compressible inviscid fluids 3:165
 - dissipative fluids 2:198
 - Hamiltonian fluid dynamics 2:593–600
 - Newtonian fluids 3:492–502
 - non-Newtonian fluids 3:560–567
 - perfect fluid equations 2:196, 3:160
 - shear-dependent viscosity fluids 3:563
 - spinning fluids 2:193
 - superfluids 1:316, 5:115–121, 5:395
 - for turbulence theories 5:298
 - viscosity fluids 3:563
 - viscous incompressible fluids 5:369–379
 - yield stress fluids 3:565
- Fluid flows
 - classical instabilities 5:2
 - equations of motion 5:1
 - Euler equation 5:5
 - geometry and topology effects 5:274
 - inviscid compressible fluids 1:596
 - multicomponent fluid flow interfaces 3:135–144
 - Navier–Stokes equations 5:3
 - overview 5:1–7
 - viscous compressible fluids 1:602
- Fluid mechanics
 - advection–diffusion 2:369
 - elliptic equations 2:366
 - finite volume principles 2:370
 - hyperbolic equations 2:367
 - interpolation 2:368
 - and knots 5:274
 - meshes 2:368
 - models 2:365
 - Navier–Stokes projection methods 2:373
 - overview 2:365–374
 - parabolic equations 2:367
 - Poisson problem 2:371
 - Stokes equations 2:367, 2:370, 2:372
 - transport equation 2:371

- Fluid patterns, topological changes,
dissipation, reconnection 5:277
- Fluid sources, Einstein equations 2:196
- Flux vacua
in IIB string theory 4:326
overview 4:323–329
- Focal conic domains, in smectics 3:327
- Fock space
and bosons and fermions 1:320–321
coherent states 1:539–540
- Formal deformation quantization,
manifolds 2:19
- Forward problems, wave propagation
controllability 1:341
dynamical system 1:340
Laplacian on waves 1:341
manifold 1:340
spectral problem 1:342
- Forward scattering, and dispersion
relations 2:96
- Forward two-particle scattering amplitude,
dispersion relation 2:88–89
- Four-body problems, classical gravitational
many-body problems 2:579
- Four dimensions
even theory two-particle bound
states 4:486
spin foams 4:651
- \hbar -Fourier integral operators
and classical mechanics 2:707
overview 2:708
- Fourier–Mukai transform
and algebraically completely
integrable 3:77
for elliptic fibrations 2:380
overview 2:379–386
and spectral covers 2:383
transforms 2:379
vector bundles for heterotic
strings 2:382
- Fourier transform
compact groups, noncommutative
Fourier transform 1:583
continuous q -ultraspherical polynomial
as 4:110
for Feynman path integrals 2:309
functional analysis overview 1:91
Haar measure and invariant
integral 1:582
heat conduction in gases 2:375
heat conduction in insulating
crystals 2:376
and macroscopic law 2:374
nonlinear Fourier transform 3:97
orthogonality relations 1:582
overview 2:374–379
Peter–Weyl theorem 1:582
for separation of variables 4:526
statistical mechanics paradigm 2:376
and turbulence 5:411
- Four-manifolds
with boundary 2:393
computations and structure
theorems 2:459
Donaldson’s polynomial
invariants 2:389
exotic structures, and Seiberg–Witten
theory 2:392
four dimensions case 2:388
in gauge theory 2:475, 2:457–463
instantons and Donaldson
invariants 2:457
manifolds with boundary 2:462
monopoles and Seiberg–Witten
invariants 2:461
open, exotic structures on 2:392
overview 2:386–394
PL *vs.* DIFF 2:387
and Seiberg–Witten equations 2:391
structures for 2:386
symplectic and Kähler four-
manifolds 2:393
and topological QFT 2:390
TOP *vs.* PL 2:387
- FPU phenomenon *see* Fermi–Pasta–Ulam
phenomena
- FQHE *see* Fractional quantum
Hall effect
- Fractal attractors, and nonequilibrium
statistical mechanics 3:537
- Fractal dimensions
box dimensions 2:395
dissipative dynamical systems 2:106
embedology 2:400
future work 2:401
gauge functions 2:395
geometric constructions 2:396
Hausdorff dimension 2:395
Hénon attractor 2:399
hyperbolic measures 2:397
infinite-dimensional dynamical
systems 2:400
Julia and Mandelbrot sets 2:400
local Lyapunov dimension 2:398
logistic map 2:398
Lorenz attractor 2:399
multifractal analysis 2:397
overview 2:394–402
probability 2:401
Smale horseshoe 2:398
spiral trajectories 2:400
thermodynamic formalism 2:396
- Fractals
anomalous diffusion 3:480
self-similar, and geometric measure
theory 2:521
- Fractional charge, and fractional quantum
Hall effect 2:409
- Fractional Dirichlet branes, D3, orbifold
and super-QCD 2:464
- Fractional quantum Hall effect
bilayers 2:410
composite fermions 2:406, 2:410
consequences 2:408
edge states 2:410
fermion Chern–Simons theory 2:407
Hall effect overview 2:403
IQHE comparison 4:246
Landau levels 2:404
lowest Landau level problem 2:406
overview 2:402–411, 4:250
pairing 2:410
plateau origins 2:405
spin 2:410
Tsui–Stormer–Gossard effect 2:404,
2:410
von Klitzing effect overview 2:403
- Frames
classical mechanics overview 1:11–12
Clifford algebras 1:520
invariance and objectivity, and non-
Newtonian fluids 3:563
vector bundles 1:488
- Fredholm alternative, second-order elliptic
problems 2:224
- Fredholm determinant
operators and line bundles 4:318
properties 4:316
- Fredholm mappings, and complex
analysis 4:546
- Fredholm operators
functional analysis overview 1:92
and line bundles 4:318
- Free convolution, operations 2:419
- Free Dirac field, properties 2:78
- Free discontinuity
and gamma-convergence 2:455
overview 2:411–417
- Freedom
in constructive QFT 1:623
and perturbative quantum
chromodynamics 4:149
quantum chromodynamics 4:147
- Free energy
Parisi representation, spin glasses 4:662
partition function 3:270
and Pirogov–Sinai theory 4:62
prepotential 3:271
- Free entropy, and free probability
theory 2:422
- Free Fermi gas, physical observables 2:301
- Free field theory
and Euclidean field theory 2:260
operator product expansion 3:617
regularized free field 1:618
as thermal field model 5:233
zero-rest-mass free fields 5:313
- Free gauge fields, quantization 4:43
- Free independence, with amalgamation
over subalgebra 2:418

- Free interfaces, overview 2:411–417
 - Freeness
 - noncommutative random variable 2:417
 - random matrices 2:421
 - Free probability theory
 - asymptotic freeness of random matrices 2:421
 - convolution operations 2:419
 - free entropy 2:422
 - free independence with amalgamation 2:418
 - noncommutative random variable freeness 2:417
 - noncrossing partitions 2:420
 - overview 2:417–424
 - semicircle law 2:419
 - Free rigid body, reconstruction
 - phases 1:664
 - Free rotators, classical mechanics
 - overview 1:9
 - Free wheels, classical mechanics
 - overview 1:9
 - Freezing phenomenology
 - spin glasses 2:553
 - structural glasses 2:554
 - Frequencies, and Hamiltonian system
 - stability 2:631
 - Frequency locking, in periodic orbit
 - bifurcations 1:286
 - Frequency set, h -pseudodifferential operator application 2:709
 - Friedmann–Lemaître models
 - approximate solutions 1:655
 - and mathematical cosmology 1:654
 - Friedman–Robertson–Walker metric
 - FRW–TOV shock matching 4:565
 - shock wave refinement
 - overview 4:559–570
 - Frobenius manifolds, and WDVV
 - equations
 - canonical coordinates 5:441
 - deformed flat connection 5:439
 - examples 5:444
 - and integrable systems 5:446
 - intersection form 5:440
 - moduli 5:442
 - overview 5:438–447
 - Front-tracking methods, for
 - multicomponent fluid flow 3:137
 - FRW models *see* Friedman–Robertson–Walker metric
 - FTIs *see* Finite-type invariants
 - Fuchsian equations, and 2D Ising
 - model 5:326
 - Fuchsian linear systems, on Riemann
 - sphere 4:436
 - Fully nonlinear evolution equations,
 - parabolic type 2:271
 - Fully packed loop configurations, in
 - combinatorics 1:570
 - Functional analysis
 - duality 1:90
 - Fourier transform 1:91
 - Fredholm operators 1:92
 - one-parameter groups 1:93
 - operator algebras 1:95
 - overview 1:88–96
 - spectral theory 1:94
 - topological vector spaces 1:89
 - Functional equations
 - Baxter’s T – Q relation 2:161
 - hard-hexagon model 2:161
 - Functional equations, for integrable
 - systems
 - algebraic ansatz 2:427
 - Baker–Akiezer functions 2:431
 - geometric ansatz 2:430
 - ground-state factorization 2:431
 - Lax pairs 2:425
 - many-body problems 3:109
 - overview 2:425–433
 - Poincaré algebra 2:428
 - quantum Calogero–Moser systems 2:428
 - quantum Ruijsenaars–Schneider models 2:429
 - topology 2:432
 - Functional integration
 - basic techniques 2:434
 - Chern–Simons functional integrals 1:497
 - coarse-graining operator 2:446
 - degenerate Hessians 2:442
 - domain 2:434
 - Gaussian scaling 2:445
 - Gaussian volume elements 2:436
 - homotopy theorem 2:444
 - n -indistinguishable particles 2:444
 - instantons 2:445
 - Koszul formula 2:439
 - as multipurpose tool 2:444
 - non-Gaussian volume elements 2:438
 - operator formalism 2:440
 - overview 2:434–447
 - path integrals 2:435
 - in QFT 2:446
 - semiclassical expansions 2:441
 - spinning top 2:445
 - Witten’s functional integral 3:225
 - WKB approximations 2:442
 - Functionals
 - action functional 2:642
 - convex functional 2:414
 - energy functional 5:355
 - and Ginzburg–Landau energies 5:360
 - nonconvex functional 2:415
 - in topological QFT 1:232
 - Yang–Mills functional 2:469
 - Functional spaces
 - cost function minimization over 3:4
 - examples 3:4
 - topologies 1:135
 - η -Function identities, and root
 - systems 4:112
 - Fusion rules, and conformal field
 - theory 5:321
 - Fyodorov’s method, random matrix theory
 - supersymmetry method 5:154
- ## G
- Galactic microlensing, as
 - phenomenon 2:571
 - Galerkin approximation, and viscous
 - incompressible fluids
 - classical regularity 5:377
 - initial values assumption 5:378
 - overview 5:376
 - passage to limit 5:377
 - and pressure 5:377
 - Galerkin method, in fluid mechanics
 - approximations 2:369–370
 - Galilean invariance, limits 1:45
 - Galilean relativity, inconsistency 1:44
 - Galois theory
 - and conformal field theory 3:604
 - and Hamiltonian system
 - nonintegrability 2:629
 - overview 3:602
 - and renormalization 3:606
 - Galvanism, electromagnetism
 - overview 1:50
 - Gamma-convergence, and homogenization
 - definition and properties 2:449
 - dimension reduction problems 2:453
 - free-discontinuity problems 2:455
 - homogenization problems 2:450
 - overview 2:449–457
 - perforated domains 2:452
 - phase transition problems 2:454
 - Gamma function, characteristics 3:637
 - q -Gamma functions, in q -hypergeometric
 - series 4:107
 - Gamma volume elements, and functional
 - integration 2:439
 - Gases
 - Bose–Einstein condensates 1:313
 - heat conduction 2:375
 - Gauge conditions, and constrained
 - systems 1:614
 - Gauged Poisson structures, and cotangent
 - bundle reduction 1:665
 - Gauged sigma model, and Mathai–Quillen
 - formalism 3:395
 - Gauge fields
 - and perturbation theory 4:35
 - quantization 4:43
 - Gauge fixing
 - and abelian BF theories 1:260
 - and Batalin–Vilkovisky
 - quantization 1:250
 - and symmetry breaking 5:202

- Gauge functions, in dynamics 2:395
- Gauge groups
 - and A–B flat connections 1:196
 - on branes 1:360
 - with differential forms 1:148
 - extensons, and gerbes 2:545
 - 3D Chern–Simons theory with 4:495
- Gauge invariance
 - in constrained systems 1:611–612
 - differential forms 1:145
 - in electric–magnetic duality 2:202
 - nonabelian BF theories 1:258
 - and noncommutative gauge theory 3:519
 - and quantum chromodynamics 4:146
- Gauge models, and BRST quantization 1:391
- Gauge potential
 - differential forms 1:145
 - and field strength 1:147
- Gauge symmetries
 - and first-class constraints 1:614
 - and quantum spin system symmetry 4:299
- Gauge theory
 - ADHM construction 2:472
 - anomaly cancellation 1:207
 - applications 2:479
 - and brane construction, overview 1:360–367
 - and BRST quantization 1:391
 - Chern–Simons gauge theory 5:284, 5:288
 - computations and structure theorems 2:459
 - with differential forms 1:141–150
 - Dirac fields overview 2:67–74
 - in four dimensions 3:569
 - four-manifold invariant overview 2:457–463
 - fractional D3 branes 2:464
 - gluing techniques 2:471
 - higher dimensions 2:471
 - hyper-Kähler quotients 2:475
 - as infinite-dimensional Lie group example 3:40
 - instantons and Donaldson invariants 2:457
 - instantons in four dimensions 2:470
 - integrable systems 2:473
 - Jones–Witten theory 2:478
 - lattices 1:211, 3:275–281
 - low-dimensional topology 2:475
 - manifolds with boundary 2:462
 - Mathai–Quillen formalism applications 3:398
 - moduli spaces 2:470
 - moduli space topology 2:474
 - moment maps 2:473
 - monopoles and Seiberg–Witten invariants 2:461
 - noncommutative gauge theory 3:519
 - overview 2:468–481, 3:568–578, 4:236–244
 - perturbative construction 4:45
 - perturbative quantization problems 4:41
 - renormalization and symmetries 4:404
 - and Seiberg dualities 4:503
 - and Seiberg–Witten duality 4:507
 - Seiberg–Witten invariants 2:478
 - spontaneously broken types 5:201
 - from strings, overview 2:463–468
 - and symmetries 5:167
 - thermal gauge theory 1:176
 - in three dimensions 3:576
 - transformations 2:468
 - in two dimensions 3:575
 - Uhlenbeck’s theorem 2:469
 - Ward correspondence 2:472
 - wrapped D5 branes 2:466
 - Yang–Mills functional 2:469
 - Yang–Mills theory 3:570
- Gauge transformations
 - and Batalin–Vilkovisky quantization 1:247
 - electromagnetism overview 1:50
 - overview 2:468
 - and quantum group bundles 4:242
- Gauss–Bonnet theorem, overview 3:27
- Gauss hypergeometry, characteristics 3:640
- q -Gauss hypergeometry
 - q -analogue of Euler’s integral representation 4:107
 - evaluation formulas in special points 4:108
 - general transformation formulas 4:108
 - second order q -difference equation 4:108
 - terminating case transformation formulas 4:108
- Gaussian domination
 - in classical systems 4:380
 - in quantum systems, and phase transitions 4:384
- Gaussian ensembles, and random matrix theory supersymmetry method 5:151
- Gaussian integrals, determinants as 5:153
- Gaussian measures
 - and entanglement 2:232
 - and renormalization groups 2:274
- Gaussian orthogonal ensembles, in random matrix theory 4:339
- Gaussian path integrals, and abelian BF theories 1:260
- Gaussian processes, spin glasses 3:408
- Gaussian scaling, and functional integration 2:445
- Gaussian states, cloning 3:631
- Gaussian superintegrals, random matrix theory supersymmetry method 5:157
- Gaussian symplectic ensembles, in random matrix theory 4:339
- Gaussian unitary ensembles, in random matrix theory 4:339
- Gaussian volume elements, and functional integration 2:436
- Gauss map, and minimal submanifolds 3:421
- Gauss sums, and finite-type invariants 2:346
- Gauss theorem, and stationary charge distributions 1:43
- Gazeau–Klauder coherent states, definition 1:541
- Gelfand–Naimark spectrum, and C^* -algebras 1:393
- Gelfand’s problem, and Iljin–Oleinik theorem 1:447
- General elliptic problems, and linear theory 2:223
- General index theorem, overview 3:29
- Generalized coherent states, characteristics 1:544
- Generalized cohomology theories, examples 1:551
- Generalized knots, and nonabelian BF theories 1:264
- Generalized Kostka numbers, Macdonald polynomials 4:114
- Generalized Mathisson–Papapetrou equation of motion, and Einstein–Cartan theory 2:193
- Generalized synchronization, chaotic systems 5:221
- General relativity theory
 - basic concepts and laws 3:503
 - binary pulsar tests 2:486
 - black holes 2:512
 - in canonical form 3:339
 - canonical general relativity 1:412–417
 - Cauchy problem 1:609
 - conformal flow of metrics 2:516
 - constraint equations 1:608
 - conventions and units 1:604
 - and Einstein–Cartan theory 2:194
 - and Einstein equations 2:67, 2:492
 - Einstein equations boundary conditions 1:610
 - Einstein equivalence principle 2:482
 - and energy conservation 2:493
 - equivalence principle 2:487
 - evolution equations 1:609
 - experimental tests overview 2:481–487
 - geometric analysis overview 2:502–509
 - gravitational attraction 2:511
 - gravitational energy 2:512
 - gravity as curved spacetime 2:489
 - inequalities overview 2:510–520

- inverse mean curvature flow 2:515
- local energy density 2:511
- metric tensor interpretation 2:491
- and Newton's laws applications and problems 3:508
- and Newton's laws limits 3:507
- origin of difficulties 1:412
- overview 1:604–610, 2:487–494, 3:503–509
- positive mass theorem 2:512
- Riemann curvature and geodesic deviation 2:491
- Riemannian Penrose inequality problem 2:514, 2:519
- Schwarzschild metric 2:513
- solar-system tests 2:484, 2:485
- spinor algebra 4:668
- spinor calculus 4:669
- superharmonic functions 2:512
- and symmetries 5:171
- 3 + 1 spacetime split 1:606
- tidal effect 2:488
- topological reasoning 4:617
- total mass 2:510
- General three-body problem, as classical gravitational problem 2:577
- Generating functions
 - asymptotic information extraction 1:560
 - in combinatorics 1:556
 - continuous q -ultraspherical polynomials 4:111
 - Donaldson–Witten function 2:113
 - equation solution 1:559
- C^1 -Generic diffeomorphisms
 - dominated splittings *vs.* wild behavior 2:501
 - hyperbolic properties 2:501
- C^1 -Generic dynamical systems
 - chain recurrence classes–homoclinic classes 2:500
 - conservative systems 2:501
 - ergodic properties 2:500
 - periodic orbit densities 2:499
 - periodic orbit perturbations 2:499
 - tame and wild systems 2:500
- C^r -Generic dynamical systems
 - Kupka–Smale theorem 2:498
 - low-dimensional systems 2:498
 - surface diffeomorphism wild behavior 2:498
- Genericity, in optical caustics 3:625
- C^2 -Genericity, surface diffeomorphism wild behavior 2:498
- Generic nonintegrability, classical mechanics overview 1:15
- Generic systems, and Lyapunov exponents 3:354
- Geodesic deviation, and Riemann curvature 2:491
- Geodesic flows
 - as hyperbolic dynamical system example 2:724
 - and quantum ergodicity 4:189
 - and wave group 4:184
- Geodesic incompleteness, and Lorentzian splitting theorem 3:346
- Geodesics
 - and arithmetic quantum chaos 1:216
 - on compact manifolds 3:436
 - and 2-step groups 4:98
- Geometrical optics, and wave equation 5:406
- Geometric analysis, and general relativity
 - center of mass 2:507
 - Jang's equation 2:506
 - marginally trapped surfaces 2:505
 - maximal and related surfaces 2:507
 - minimal and related surfaces 2:503
 - minimal surfaces and positive mass 2:504
 - null hypersurfaces 2:508
 - overview 2:502–509
- Geometric integrability scheme, quadrilateral lattice 3:79
- Geometric measure theory
 - area formula 2:523
 - currents, boundary, mass 2:524
 - currents with finite mass 2:525
 - current solutions and tools 2:527
 - currents and plateau problem 2:526
 - differential forms and Stokes theorem 2:524
 - Hausdorff dimension 2:521
 - Hausdorff measure 2:521
 - integral currents 2:525
 - k -covectors 2:524
 - k -vectors and exterior product 2:523
 - normal currents 2:525
 - overview 2:520–528
 - rectifiable currents 2:525
 - rectifiable sets 2:522
 - rectifiable sets with finite measure 2:522
 - self-similar fractals 2:521
 - simple vectors and orientation 2:523
- Geometric modular action, and Tomita–Takesaki modular theory applications 5:256
- Geometric operators, in quantum geometry 4:232
- Geometric optics, and contact manifolds 1:635
- Geometric phases
 - adiabatic evolution 2:530
 - Aharonov–Anandan phase 2:532
 - Berry's phase 2:531
 - generalization 2:533
 - overview 2:528–534
 - Pancharatnam connection 2:529
 - quantum jumps 2:530
 - space of rays 2:528
- Geometric series, example 3:639
- Geometric theory
 - domains in C^n 4:540
 - and Stein manifolds 4:543
- Geometric transitions, and large- N duality 3:267
- Geometry concepts
 - algebraic geometry 3:65–78, 3:81, 4:548
 - approach to supermanifolds 5:130
 - arithmetic algebraic geometry 3:604
 - calibrated geometry 1:398–402
 - classical differential geometry 3:78
 - complex geometry 1:634
 - confluent hypergeometry 3:640
 - constrained systems 1:616
 - constructions for fractal analysis 2:396
 - differential geometry 1:33–40
 - elliptic hypergeometric series 4:115
 - enumerative geometry 3:154
 - Feigenbaum Julia sets, and renormalization 5:347
 - and finite group symmetry breaking 2:324
 - in fluid flows 5:274
 - gauge theories from branes 1:363
 - Gauss hypergeometry 3:640
 - and holomorphic dynamics 2:658
 - Lorentzian geometry 3:343–349
 - mapping theory 4:547
 - Metric-affine geometry 2:192
 - Minkowskian geometry 2:94
 - Montel's theorem and Fredholm mappings 4:546
 - noncommutative geometry 2:699, 3:509–515, 3:515–524, 4:8–12
 - orbit space, and finite group symmetry breaking 2:325
 - and perfect lens 3:490
 - phase coexistence 1:74
 - quantum geometry 4:230–236
 - resonance geometry 2:632
 - Riemann–Cartan geometry 2:192
 - Riemannian geometry 1:634
 - symplectic and Poisson groupoids 3:314
 - symplectic geometry 1:633
 - transitions 3:273
 - twistors 5:311
 - twistor theory applications 5:307
 - 2D conformal geometry 5:317–322
- Geophysical dynamics
 - atmospheric PEs 2:537
 - ocean PEs 2:535
 - overview 2:534–539
- Gerbes, in QFT
 - from canonical quantization 2:541
 - definitions and example 2:539
 - and gauge group extensions 2:545
 - index theory 2:544
 - overview 2:539–546

- Geroch–Kronheimer–Penrose formulation, and spacetime causal boundary 1:329
- Gerstenhaber algebras, operads in 3:613
- Ghirlanda–Guerra relations, mean-field spin glasses 3:412
- Giant luminous arcs, and gravitational lensing 2:574
- Gibbsian properties, determinantal random fields 2:52
- Gibbs measures
and dimer problems 2:61
spin glass models 3:409
- Gibbs states
and interfaces 5:57
large deviations properties 3:261
- G-invariant polynomials, and finite group symmetry breaking 2:324
- Ginzburg–Landau energy, variational techniques for
functional with magnetic field 5:360
issues and methods 5:357
overview 5:355–362
- Ginzburg–Landau equations
and applied magnetic field solutions 2:549
and finite group symmetry breaking 2:326–327
in fluid dynamics 1:284
and nonlinear Schrödinger equations 3:552
overview 2:547–552
persistent current solutions 2:549
simplified model 2:551
time-dependent equations 2:551
in whole space 2:548
- Ginzburg–Landau model
energy functional 5:355
interacting stochastic particle systems 3:131–132
overview 5:355
reductions 5:356
simplified model 5:357
vortices and critical fields 5:357
- GKP formulation *see*
Geroch–Kronheimer–Penrose formulation
- Glashow–Weinberg–Salam theory
see also Electroweak theory
 B and L nonconservation 2:213
CKM quark mixing 2:212
CP violation 2:212
Fermi interactions 2:211
global fit 2:213
masses 2:211
neutral currents 2:211
overview 2:209, 2:210
- Glassy disordered systems
aging and slow dynamics 2:555
mean-field models 2:556
overview 2:553–560
spin glasses 2:553
structural glasses 2:554
- Glassy dynamics, interacting stochastic particle systems 3:132
- Glauber dynamics, and interacting stochastic particle systems 3:131
- Glimm classification, C^* -algebras 1:394–395
- Global analyticity properties, and QFT analytic program 4:471
- Global asymptotics, and caustics 3:625
- Global attractors, dissipative dynamical systems
dimension 2:105
generalization 2:108
overview 2:104
structure 2:106
- Global bifurcation theory, overview 1:280
- Global classes, and b -pseudodifferential calculus 2:705
- Global equilibrium, in black hole mechanics 1:300
- Global fit, and Glashow–Weinberg–Salam theory 2:213
- Global large waves, semilinear wave equations 4:522
- Global shape functions, in fluid mechanics 2:368–369
- Global small waves, semilinear wave equations 4:523
- Global symmetries, breaking 5:198
- Global wavelet spectrum, and turbulence 5:411
- GL_2 -system, and class field theory 3:603
- Gluings techniques, and gauge theory 2:471
- GMS solitons, and noncommutative geometry from strings 3:521
- GNS construction, C^* -algebras 4:91
- GOE *see* Gaussian orthogonal ensembles
- Goldfish model, attributes 3:111
- Goldstone bosons, and symmetry breaking 5:200
- Goldstone theorem, and quantum fields 4:225
- Goussarov–Habiro filtrations, finite-type invariants 2:350
- GPB *see* Gravity Probe B
- GP equation *see* Gross–Pitaevskii equation
- Graded algebras
Clifford algebras 1:522, 1:523
definition 2:560
- Graded Lie algebras, definition 2:560
- Graded Poisson algebras
and AKSZ 2:566
algebraic topology examples 2:564
BRST quantization in Hamiltonian formalism 2:564
BV quantization in Lagrangian formalism 2:566
from cohomology 2:566
- definitions 2:560
- graded symplectic manifolds 2:564
- Hochschild cohomology 2:563
- Lie algebroid cohomology 2:563
- Lie algebroid example 2:562
- Lie–Rinehart algebras 2:563
- Lie–Rinehart cohomology 2:563
- overview 2:560–567
- Schouten–Nijenhuis bracket
example 2:562
- shifted cotangent bundles 2:564
- Graded spectral sequences,
convergence 4:626
- Graded symplectic manifolds,
overview 2:564
- Graded vector spaces, definition 2:560
- Gradient percolation, characteristics 4:26
- Grand-canonical ensemble, in
Falicov–Kimball model
mathematics 2:285
- Grand-canonical phase diagram,
Falicov–Kimball model 2:287
- Graph manifolds, and Hamiltonian reduction of Einstein’s equations 2:609
- Graphs
automorphisms, and quantum spin system symmetry 4:299
in combinatorics 2:62
for finite Weyl system
representation 2:332
Lie and Hopf algebras 2:680
minimal, and minimal submanifolds 3:421
nonamenable, percolation on 4:26
in perturbation theory 2:56
rigid vertex, and Vassiliev invariants 3:223
Young graphs 4:351
- Gravitational attraction, in general relativity 2:511
- Gravitational buoyancy convection,
Newtonian fluids 3:498
- Gravitational collapse, critical phenomena
and cosmic censorship 1:672
dynamical systems picture 1:669
overview 1:668–673
statistical mechanical analogy 1:671
and universality 1:672
- Gravitational energy
capillary surfaces 1:434–435
in general relativity 2:512
- Gravitational lensing
basics 2:568
cosmic shear 2:575
critical surface mass density 2:569
Einstein radius 2:569
Einstein rings 2:574
Fermat’s theorem 2:571
galactic microlensing 2:571

- giant luminous arcs and arclets 2:574
 - image positions 2:569
 - lens equation 2:568
 - lens mapping 2:570
 - magnifications 2:569
 - multiply-imaged quasars 2:572
 - nonsingular isothermal sphere 2:570
 - overview 2:567–575
 - quasar microlensing 2:573
 - singular isothermal sphere 2:570
 - statistical lensing 2:575
 - time delay 2:571
 - weak lensing 2:575
 - Gravitational many-body problems, classical *see* Classical gravitational many-body problems
 - Gravitational waves
 - general relativity tests 2:486
 - overview 2:582–586
 - Gravitons, nonlinear, and twistor theory 5:305
 - Gravity
 - A–B effects 1:198
 - cogravity 1:265
 - as curved spacetime 2:489
 - Dirac fields, overview 2:67–74
 - Einstein–Cartan theory
 - Bianchi identities 2:193
 - generalized Mathisson–Papapetrou equation of motion 2:193
 - and general relativity theory 2:194
 - identity from local invariance 2:192
 - projective transformations 2:192
 - Sciama–Kibble field equations 2:193
 - spinning fluid 2:193
 - spin and torsion 2:194
 - Einstein gravity 5:454
 - higher-dimensional, in brane worlds 1:367
 - and nonabelian BF theories 1:263, 1:264
 - quantum gravity 3:215–220
 - supergravity 5:122–128
 - 3D 4:501
 - 2D topological gravity 5:264–270
 - Gravity Probe B, and general relativity solar system tests 2:485
 - Gravothermal oscillations, classical gravitational many-body problems 2:581
 - Gravothermal runaway, classical gravitational many-body problems 2:580–581
 - Green functions
 - and bounded point vortex motion 4:76
 - peakon solution generalizations for 4:16
 - in perturbation theory 4:31
 - and quantum chromodynamics 4:146
 - random matrix theory supersymmetry method 5:152
 - and Riemann surfaces 4:426
 - Griffiths inequalities, equilibrium statistical mechanics overview 1:69
 - Gromov–Witten invariants, and topological QFT 5:287
 - Gross–Pitaevskii equation, and Bose–Einstein condensates 1:315
 - Grothendieck–Riemann–Roch theorem, for smooth manifolds 3:247
 - Grothendieck spectral sequence, properties 4:632
 - Ground-state energy, quantum Calogero–Moser systems 4:126
 - Ground-state factorization, for integrable systems 2:431
 - Ground-state phase diagrams, and Pirogov–Sinai theory 4:60
 - Ground-state wave function, quantum Calogero–Moser systems 4:126
 - Groupoids
 - definition 3:312
 - examples 3:313
 - identification and notations 3:313
 - overview 3:312–320
 - properties 3:313
 - Growth processes, and limit laws F_β 3:104
 - Growth processes, in random matrix theory
 - airy process 2:590
 - basic model 2:587
 - limit theorems 2:589
 - overview 2:586–591
 - Robinson–Schensted–Knuth correspondence 2:588
 - variations 2:591
 - GRT *see* General relativity theory
 - GSE *see* Gaussian symplectic ensembles
 - GUE *see* Gaussian unitary ensembles
 - Guerra’s integral representation, mean-field spin glasses 3:410
 - Gupta–Bleuler gauge procedure
 - definition 3:17
 - overview 3:21
 - Gutzwiller’s formula, \hbar -pseudodifferential operator application 2:710
 - GWS theory *see* Glashow–Weinberg–Salam theory
 - Gyroscope precession, in general relativity solar system tests 2:485
- ## H
- H_0
 - construction 3:211
 - deformations 3:213
 - and Heegaard Floer homology 3:210
 - H_2 , construction 3:213
 - Haag–Ruelle theory, and scattering 4:457
 - Haar measure, compact groups 1:582
 - Habiro ring, and cyclotomic expansion 2:354
 - Hadamard condition, and \ast -algebra construction 4:204
 - Hadronic final states, and perturbative quantum chromodynamics 4:152
 - Hadrons, relativistic local current algebra 1:674
 - Hahn polynomials, characteristics 3:644
 - q -Hahn polynomials, definition 4:112
 - Hall effect, overview 2:403
 - Hall–Littlewood polynomials, Macdonald polynomials 4:114
 - Hamiltonian BRST construction, overview 1:389
 - Hamiltonian constraint, and quantum gravity 3:218
 - Hamiltonian dynamics
 - and contact manifolds 1:635
 - and quantum spin systems 4:301
 - and semigroups 4:162
 - Hamiltonian fluid dynamics
 - associated variables 2:596
 - basic principle 2:596
 - canonical description 2:598
 - Eulerian action principles 2:597
 - and Eulerian variables 2:595
 - and finite-dimensional systems 2:598
 - ideal fluid noncanonical Poisson brackets 2:599
 - and Lagrangian variables 2:595
 - into noncanonical Poisson brackets 2:598
 - noncanonical structure 2:593
 - overview 2:593–600
 - Hamiltonian formalism
 - BRST quantization in 2:564
 - equations of motion 1:3
 - and solitons 4:597
 - Hamiltonian group actions
 - convexity theorems 2:603
 - moment map properties 2:601
 - normal form theorem 2:603
 - overview 2:600–607
 - reduction in stages 2:603
 - symplectic quotient 2:602
 - symplectic quotient cohomology rings 2:605
 - and toric manifolds 2:604
 - Hamiltonian maps, and integrable discrete systems 3:60
 - Hamiltonian reduction, Einstein’s equations
 - Bianchi models 2:618
 - cosmological applications 2:622
 - field equations 2:611
 - overview 2:607–623
 - reduced Hamiltonian 2:614
 - topological background 2:607
 - Hamiltonians
 - and constrained systems 1:612
 - Dirac, and fermions 1:323

- Hamiltonians (*continued*)
 Dyson's threefold way 5:208
 for Hitchin systems 3:75
 in NMR 3:596
 nuclear spin Hamiltonian 3:596
 and Painlevé equations 4:4
 quantum Calogero–Moser systems 4:125
 reduced Hamiltonian 2:614
 in soliton theory 1:291
 spin foams from 4:647
 and von Neumann algebras 4:93
- Hamiltonian systems
 averaging and exponential stability 2:631
 bi-Hamiltonian systems 3:459
 and differential Galois theory 2:629
 in hydrodynamics 2:135
 infinite-dimensional systems 3:37–44
 and instability theory, Arnol'd mechanism 2:634
 integrability equations 2:625
 integrability overview 2:624–630
 integrable Hamiltonian system 4:577
 and isolated periodic orbits 2:626
 and mechanics 2:624
 multi-Hamiltonian systems 3:459–465
 and permanent stability 2:633
 and Poincaré theorem 2:627
 and separatrices 2:627
 and sine-Gordon equation 4:577
 and singular point analysis 2:629
 stability and instability theory, overview 2:631–636
 stability theory 4:592
 topological obstructions 2:628
 and Ziglin's theory 2:629
- Hamiltonian vector fields, and groupoids 3:315
- Hamilton–Jacobi–Bellman equation, and ordinary differential equation control 1:639
- Hamilton–Jacobi equation
 and dynamical systems, overview 2:636–644
 entropy and optimal control 3:363
 and fluctuations 3:357–358
 and H theorem 3:361
 and hydrodynamics decomposition 3:362
 and minimum dissipation principle 3:363
 and nonequilibrium fluctuation dissipation relation 3:361
 on Riemannian manifolds 4:527
 and thermodynamic fluctuations 3:360
- Hamilton's equations, and Lie–Poisson bracket 3:38
- Hamilton's equations, on Poisson manifolds
 cotangent bundles 3:38
- finite-dimensional classical mechanics 3:37
- infinite-dimensional classical field theory 3:37
- overview 3:37
- symmetry reduction 3:39
- symplectic manifolds 3:38
- Hankel functions, characteristics 3:638
- Hard-hexagon models
 commuting transfer matrices 2:158
 corner transfer matrices 2:162
 definition 2:157
 functional equation 2:161
 one-point functions 2:163
 overview 2:155–164
 parameterization 2:160
 Yang–Baxter equations 2:158
- Hard spheres, and Boltzmann equation 1:308
- Hardy–Pomeau–De Pazzis model, lattice gas automata 1:465
- Harmonic analysis, and wave equation 5:404
- Harmonic functions, on Young graph 4:351
- Harmonic oscillator
 and deformation quantization 2:5
 Hamilton's equations 3:37
- Hausdorff dimension
 in dynamics 2:395
 and geometric measure theory 2:521
 and holomorphic dynamics 2:659
- Hawking effects, and curved spacetime 4:209
- Hawking–Penrose singularity theorem, definition 3:347
- Hawking radiation, and branes 1:379
- HD *see* Hausdorff dimension
- Heat
 and equilibrium statistical mechanics 1:53
 and Newtonian fluids 3:494
 and nonequilibrium statistical mechanics 3:534
- Heat conduction
 in gases 2:375
 in insulating crystals 2:376
- Heat measures, on loop space 3:414, 3:416–417
- Heat semigroups, and linear evolution equations 2:266–267
- Heat theorem, equilibrium statistical mechanics overview 1:53
- Heat transfer, adiabatic piston infinite cylinder 1:167
- Heat transport, and nonequilibrium statistical mechanics 3:549
- Heavy-quark effective theory, overview 2:142
- Heavy top, reconstruction phases 1:664
- Hecke operators, and arithmetic quantum chaos 1:217
- Heegaard Floer homology
 characteristics 3:210
 overview 2:364
- Heinz's curvature estimate, and minimal submanifolds 3:425
- Heisenberg spin chains, and Bethe ansatz 1:256
- Helium-3, and topological defects 5:262
- Henkin–Shanin–Tumanov theorem, generalized Burgers equations solutions 1:451
- Hénon attractor, in dynamics 2:399
- Hénon family, and omoclinic phenomena 2:676
- Hénon-like attractors, and Lyapunov exponents 3:355
- Herivel–Lin action, in Hamiltonian fluid dynamics 2:597
- Hermiticity, N -point functions 1:237
- Hermite polynomials, overview 3:642
- Hermite volume elements, and functional integration 2:439
- Hermitian scalar products, Clifford algebra representations 1:527
- Hermitian symmetric manifolds, characteristics 1:510
- Hermitian symmetric spaces, and Berezin-type star products 2:31
- Hermitian–Yang–Mills equations, and mirror symmetry 3:443
- Heteroclinic cycles, and equivariant dynamics 5:186
- Heterotic strings
 Calabi–Yau compactification 1:588
 overview 5:104
 vector bundles for 2:382
- Heuristic analysis, in constructive QFT 1:623
- Hidden-variables theory
 deterministic, extension 4:269
 and quantum nonequilibrium 4:266
- Higgs Lagrangian, electroweak theory 5:36
- Higgs model
 critical coupling 1:156
 interaction energy 1:153
 overview 1:151
 radially symmetric vortices 1:152
 research developments 1:159
 useful abbreviations 1:159
 vortex scattering 1:155
- Higgs potential, minimal supersymmetric standard model 5:142
- Higgs vortices
 characteristics 4:605
 overview 1:151–160
- Higher Chow groups, and intersection theory 3:155

- Higher-dimensional compactification, in
 superstring theory 1:590
- Higher-dimensional gravity, in brane
 worlds 1:367
- High-temperature superconductivity
 antiferromagnetism unification via $SO(5)$
 theory 2:648
- Hubbard model symmetry 2:647
- overview 2:645–652
- Hilbert space
 and bosons and fermions 1:319
 and functional integration 2:440
 and IMQFT 4:216
 and indefinite metrics 3:21
 and linear evolution
 equations 2:266–267
 and quantum spin system
 symmetry 4:298
- Hill's problem, classical gravitational
 many-body problems 2:577
- Hirota method
 and multidimensional quadrilateral
 lattices 3:83
 and solitons 4:597
- Hirzebruch signature theorem,
 overview 3:28
- Hitchin abelianization program,
 characteristics 3:75
- Hitchin map, properties 3:75
- Hitchin systems
 characteristics 3:74
 explicit Hamiltonians for 3:75
- HJB equation *see*
 Hamilton–Jacobi–Bellman equation
- HJE *see* Hamilton–Jacobi equation
- Hochschild cochain complex,
 cohomology 2:563
- Hochschild cohomology, and Hopf
 algebras 2:684
- Hodge decomposition, cohomology
 theories 1:550
- Hodge duals, Einstein–Cartan theory 2:191
- Hölderian functions, multifractal
 analysis 2:397
- Holevo–Schumacher–Westmoreland
 random codes, and quantum channel
 capacity 1:428
- Holography
 and anti-de Sitter spacetimes–conformal
 field theory 5:139
 in Arakelov geometry 3:605
 and black hole statistical
 mechanics 1:381
- Holomorphic dynamics
 and area 2:658
 Böttcher coordinate 2:656
 and combinatorial equivalence 2:656
 and conformal measure 2:659
 connectivity 2:658
 examples 2:654
- Fatou and Julia sets 2:654
 and Fatou set 2:655
 fundamental conjectures 2:657
 and Hausdorff dimension 2:659
 hyperbolic maps 2:656
 Julia set properties 2:655
 local theory 2:653
 and Mandelbrot set 2:657
 and maximal entropy measure 2:659
 no wandering domains theorem
 2:655
 overview 2:652–659
 and parabolic implosion 2:658
 periodic points 2:654
 postcritically finite maps 2:656
 rational function spaces 2:656
 and structural stability 2:657
- Holomorphic motions, and structural
 stability 2:657
- Holonomic quantum fields
 and deformation theory 2:661
 overview 2:660–664
 Riemann–Hilbert problem 2:663
 2D Ising model 2:660
- Holonomy
 and A–B geometry 1:196
 Einstein manifolds 2:184
 hyper-Kähler manifolds 2:185
 quaternionic Kähler manifolds 2:186
- Homeomorphisms, circle
 finite type 2:665
 overview 2:665–672
- Homoclinic bifurcation, and singularity
 theory 4:592
- Homoclinic classes
 C^1 -generic dynamical systems 2:500
 and hyperbolicity 2:497
- Homoclinic phenomena
 Hénon family 2:676
 orbits in diffeomorphisms 2:673
 overview 2:672–678
 and positive topological entropy
 systems 2:676
 Smale horseshoe map 2:674
 tangencies 2:677
- Homogeneity, in Falicov–Kimball
 model 2:288
- Homogeneous spaces,
 overview 1:500–511
- Homogenization, and gamma-convergence
 definition and properties 2:449
 dimension reduction problems 2:453
 free-discontinuity problems 2:455
 homogenization problems 2:450
 overview 2:449–457
 perforated domains 2:452
 phase transition problems 2:454
- Homographic solutions, classical
 gravitational many-body
 problems 2:577
- Homological algebra, in BRST
 construction 1:386
- Homological mirror symmetry
 and D-branes 2:384
 Kontsevich's conjecture 3:442
- Homology, and chain complexes 1:546
- Homomorphism, induced,
 C^* -algebras 3:250
- Homotopy
 C^* -algebras 3:250
 invariance and singular homology 1:547
 in K -theory 3:247
 in multiply connected space 2:444
 operads in 3:610
 and twisted K -theory 3:252–253
 vector bundles 1:490
- Homotopy groups
 chain complex relationships 1:547
 for topological defects 5:258
 topologies 1:135
- Homotopy links, finite-type
 invariants 2:347
- Hopf algebras
 bicrossproduct Hopf algebras 1:265–274
 in braided categories 1:355
 chord diagrams 3:232
 classification 4:239
 covariant differential calculi 4:238
 higher-order bicovariant calculi 4:238
 preparations 4:238
 and quantum groups 2:687
 and renormalizable QFT
 graphs 2:680
 and Hochschild cohomology 2:684
 overview 2:678–686
 physical parameter
 diffeomorphisms 2:684
 rooted trees 2:685
 Woronowicz theorems 4:238
- Hopf bifurcation, and mode
 interactions 5:189
- Hopf–Rinow theorem, and Lorentzian
 geometry 3:343
- Horseshoes, as hyperbolic dynamical
 system example 2:724
- \hbar -Pseudodifferential calculus
 functional calculus 2:706
 overview 2:704
 self-adjointness and
 semiboundedness 2:705
- \hbar -Pseudodifferential operators
 and classical mechanics 2:701
 and Egorov's theorem 2:709
 and eigenvalue clustering 2:710
 evolution operators 2:707
 \hbar -Fourier integral operators 2:707
 frequency set application 2:709
 and Gutzwiller's formula 2:710
 overview 2:701–712
 and Poisson relation 2:710

- h -Pseudodifferential operators (*continued*)
 and quantum evolution 2:708
 and quantum mechanics 2:701
 and semiclassical mechanics 2:702
 HPP model *see* Hardy–Pomeau–De Pazzis model
 HQET *see* Heavy-quark effective theory
 HSW random codes *see*
 Holevo–Schumacher–Westmoreland random codes
 H -theorem
 and Hamilton–Jacobi equation 3:361
 and quantum equilibrium 4:267
 HTSC *see* High-temperature superconductivity
 Hubbard model
 ferromagnetism 2:715
 half-filled systems 2:714
 1D model 2:714
 overview 2:712–716
 superconductivity, symmetry 2:647
 wave–particle dualism 2:713
 Hubble constant, multiply-imaged quasars 2:573
 Hubble length,
 Friedman–Robertson–Walker metric 4:563
 Huygens–Fresnel principle, and caustics 3:624
 Huygens’ principle, and scattering theory 4:462
 Hybrid methods, for multicomponent fluid flow 3:140
 Hydrodynamics
 asymptotical models 2:133
 and Boltzmann equation 1:309
 decomposition 3:362
 Hamiltonian systems in 2:135
 and interacting particle systems
 entropy method 3:124
 hyperbolic equations 3:127
 nongradient models 3:126
 overview 3:123–129
 relative entropy method 3:128
 and interacting stochastic particle systems 3:134
 in magnetohydrodynamics
 description 3:376
 Hydrogen-like atoms, quantum electrodynamics
 muonium 4:173
 overview 4:171
 positronium 4:173
 Hydrostatic state, Newtonian fluids 3:494, 3:497
 HYM equations *see*
 Hermitian–Yang–Mills equations
 Hyperbolic billiards
 design 2:718
 in dimension 2 2:719
 in higher dimensions 2:720
 Jacobi fields and monotonicity 2:717
 overview 2:716–721
 soft billiards 2:720
 wave fronts and monotonicity 2:718
 Hyperbolic dynamical systems
 Anosov diffeomorphisms 2:724
 chains and shadowing 2:724
 dominated splitting 2:727
 and entropy formula 4:336
 for fractal analysis in dynamics 2:397
 C^1 -generic diffeomorphisms 2:501
 geodesic flows 2:724
 homoclinic classes 2:497
 horseshoes 2:724
 hyperbolic periodic orbits 2:496
 nearly integrable system
 comparison 2:126
 nonuniform hyperbolicity 2:728
 orbits and ergodic theory 2:725
 overview 2:496, 2:721–729
 partial hyperbolicity 2:727
 periodic points 2:726
 pseudohyperbolicity 2:727
 and renormalization 5:347
 smooth system division 2:721
 solenoids 2:725
 spectral decomposition 2:723
 stability 2:723
 stable ergodicity
 and transitivity 2:727
 structure and properties 2:723
 transitivity 2:723
 uniform systems 2:722
 Hyperbolic equations
 for fluid mechanics 2:367
 and interacting particle systems 3:127
 in symmetric–hyperbolic systems 5:162
 Hyperbolic maps, and Fatou’s conjecture 2:656
 Hyperbolic periodic orbits,
 and dynamical systems 2:496
 Hyperbolic spacetimes
 and $*$ -algebras 4:203
 and time machine question 4:210
 Hyperbolic surfaces, in arithmetic quantum chaos 1:213
 Hyperbolic theory, Smale’s approach 2:497
 Hyperfine splitting, hydrogen-like atoms 4:172
 q -Hypergeometric orthogonal polynomials
 Askey–Wilson polynomials 4:110
 big q -Jacobi polynomial as 4:111
 continuous q -ultraspherical polynomials 4:110
 q -Hahn polynomials 4:112
 overview 4:109
 q -Racah polynomials 4:111
 Rahman–Wilson biorthogonal rational functions 4:112
 Stieltjes–Wigert polynomials 4:112
 Hypergeometric series
 in combinatorics 1:573
 ordinary special functions 3:639
 q -Hypergeometric series
 q -beta functions 4:107
 definitions 4:105
 q -derivative 4:107
 q -gamma functions 4:107
 q -integral 4:107
 ${}_r\phi_{r-1}$ series summation and transformation formulas 4:108
 Rogers–Ramanujan identities 4:109
 special cases 4:106
 Hyper-Kähler manifolds, holonomy 2:185
 Hyper-Kähler quotients, and moduli spaces 2:475
 Hypersurfaces, compact and null 3:437
 Hyponormal operators, and spectral theory 4:641
- I**
 IBV problem *see* Initial-boundary value problem
 Icosahedral point symmetry, quasiperiodic systems 4:313
 Ideal fluid noncanonical Poisson brackets, and Hamiltonian fluid dynamics 2:599
 Ideal gases, Bose–Einstein condensates 1:313
 Identical synchronization
 bidirectional identical synchronization 5:218
 chaotic system pair
 construction 5:215
 unidirectional identical synchronization 5:217
 Iljin–Oleinik theorem, and Gelfand’s problem 1:447
 Image processing mathematics
 Bayesian framework 3:2
 denoising problem 3:3
 image restoration problem 3:1
 Markov random fields 3:3
 overview 3:1–10
 scale spaces 3:5
 variational approach 3:4
 wavelet approach 3:7
 Image restoration, basic problem 3:1
 Immanantal point processes,
 characteristics 2:53
 Impulse molecular flux density, and Newtonian fluids 3:494
 IMQFT *see* Quantum field theory with indefinite metric
 Incipient infinite cluster, and percolation theory 4:23

- Incline-plane problem, in fluid dynamics 1:282
- Incoherent matter, Einstein equations 2:196
- Incompleteness, and Lorentzian splitting theorem 3:346
- Incompressible Euler equations
 - local existence and blow-up problem 3:12
 - overview 3:10–17
 - 2D, weak solutions 3:14
- Incompressible perfect fluid
 - Cauchy problem for 3:163
 - smooth flows properties 3:161
 - stationary solutions 3:162
- Indefinite metric
 - axiomatic framework 3:20
 - Gupta–Bleuler gauge procedure 3:21
 - no-go theorems 3:18
 - overview 3:17–23
- Index media, for negative refraction 3:487
- Index theory
 - Atiyah–Singer index theorem 1:211, 3:248
 - Chern forms 3:27
 - classical Lefschetz formula 3:31
 - classic elliptic complex relating 3:26
 - de Rham elliptic complex 3:24
 - differential geometry overview 1:39
 - and Dirac operator 2:81
 - Dolbeault complex 3:28
 - Dolbeault elliptic complex 3:25
 - and equivariant cohomology 2:249
 - Euler form 3:26
 - examples and applications 3:29
 - Gauss–Bonnet theorem 3:27
 - general theorem 3:29
 - and gerbes 2:544
 - Hirzebruch signature theorem 3:28
 - Lefschetz formula 3:31
 - local boundary conditions 3:29
 - in optical caustics 3:626
 - overview 3:23–32
 - Pontrjagin forms 3:27
 - properties 3:29
 - signature elliptic complex 3:25
 - spectral boundary conditions 3:30
 - spin complex 3:28
 - spin elliptic complex 3:26
 - symmetric functionals on Banach space 3:331
 - twisted signature complex 3:25
 - Yang–Mills complex 3:28
 - Yang–Mills elliptic complex 3:25
- n -Indistinguishable particles, and functional integration 2:444
- Induced transformations, and ergodic theory 2:252
- Inductive construction, Clifford algebra representations 1:527
- Inertial manifolds, dissipative dynamical systems 2:107
- Infinite cylinder, adiabatic piston distribution 1:164
 - with finite mass 1:167
 - thermodynamic limits 1:166
- Infinite-dimensional classical field theory, and Poisson manifolds 3:37
- Infinite-dimensional dynamical systems
 - in dynamics 2:400
 - integrable PDEs as 3:118–119
- Infinite-dimensional Hamiltonian systems
 - finite- *vs.* infinite Lie groups 3:40
 - and Fourier integral operators 3:43
 - and Korteweg–de Vries equation 3:43
 - Lie groups 3:39
 - overview 3:37–44
 - and Poisson manifolds 3:37
- Infinite-dimensional Lie groups
 - Abelian gauge group 3:40
 - diffeomorphism groups 3:41
 - finite group comparison 3:40
 - loop groups 3:41
 - nonabelian gauge group 3:41
 - properties 3:39
 - subgroups of $\text{Diff}^\infty(M)$ 3:42
- Infinite-dimensional manifolds,
 - Ljusternik–Schnirelman theory 3:330
- Infinite-dimensional volume, and noncommutative geometry 4:9
- Infinite dimensions, dissipative dynamical systems
 - exponential attractors 2:107
 - fractal dimension estimates 2:106
 - global attractors 2:104, 2:108
 - global Lyapunov functions 2:106
 - inertial manifolds 2:107
 - Kolmogorov’s entropy 2:108
 - overview 2:101–110
 - spatial dynamics and chaos 2:109
 - spatio-temporal chaos 2:109
- Infinite loop spaces, and generalized cohomology 1:552
- Infinitely renormalizable unimodal maps, characteristics 5:343–344
- Infinite-particle systems, time evolution 3:540
- Infinitesimal symmetries,
 - definition 5:173–174
- Infinite Toda lattice
 - bi-Hamiltonian structure 5:237
 - Darboux–Bäcklund transformations and discretization 5:238
 - integrability 5:236
 - inverse scattering 5:237
 - model 5:236
 - multisoliton solutions 5:236
- Inflow boundary, and hyperbolic equations for fluid mechanics 2:367–368
- Information theory, classical and quantum 4:142
- Infrared behavior, in constructive QFT 1:629
- Infrared divergence, and perturbation theory 4:33
- Infrared estimates
 - in classical systems 4:381
 - and lattice system phase transitions 4:377
- Infrared inequality, equilibrium statistical mechanics overview 1:66, 1:86
- Infrared safety, and perturbative quantum chromodynamics 4:149
- Infrared–ultraviolet mixing, and noncommutative QFT 3:518
- Inhomogeneous models, quantum cosmology 4:157
- Initial-boundary value problem, nonlinear Schrödinger equation 1:349
- Initial-value formulation, Einstein equations
 - closed manifold CMC data 2:177
 - conformal method 2:176
 - conformal thin sandwich 2:178
 - gluing solutions 2:180
 - overview 2:173–182
 - parabolic methods 2:179
 - quasispherical ansatz 2:179
 - space + time field decomposition 2:174
- Injective factors, von Neumann algebras 5:383
- Injective von Neumann algebras, and nuclear C^* -algebras 4:93
- Inner products, Clifford algebras 1:519
- Insertion operads, characteristics 3:615
- Instability theory, and Hamiltonian systems
 - Arnol’d mechanism 2:634
 - overview 2:631–636
- q -Instanton, and quantum group bundles 4:243
- Instanton Floer homology, overview 2:362
- Instanton moduli space, properties 3:48
- Instanton number, Donaldson’s polynomial invariants 2:389
- Instantons
 - basic classes 2:476
 - chiral anomaly applications 1:207
 - definition 1:382
 - and Donaldson invariants 2:457
 - in four dimensions 2:470
 - and functional integration 2:445
 - in gauge theory 2:475
 - noncommutative instantons 3:522
 - in noncommutative Yang–Mills theory 3:529
 - Yang–Mills instantons 4:606
- Instanton topology
 - and ADHM construction 3:46

- Instanton topology (*continued*)
 and anti-self-dual Yang–Mills equation 3:47
 on Euclidean space 3:45
 moduli space 3:48
 overview 3:44–50
 Yang–Mills equation variational aspects 3:45
- Insulating crystals, heat conduction in 2:376
- Integer quantum Hall effect, definition 4:245
- Integrability
 in Bäcklund transformations 1:242
 bi-Hamiltonian systems 3:463
 classical mechanics overview 1:7
 and discrete systems 3:61
 Hamiltonian systems
 and differential Galois theory 2:629
 equations 2:625
 and isolated periodic orbits 2:626
 overview 2:624–630
 and Poincaré theorem 2:627
 and separatrices 2:627
 and singular point analysis 2:629
 topological obstructions 2:628
 and Ziglin’s theory 2:629
 nonlinear PDEs 3:121
 and QFT 5:338
 quantum Calogero–Moser systems 4:128
 at quantum level 3:52
- Integrable billiards, overview 1:298
- Integrable deformations, quantum Calogero–Moser systems 4:128
- Integrable discrete systems
 complete integrability 3:61
 dynamical systems 3:59
 Lagrangian and Hamiltonian formulations 3:60
 and Lax approach 3:62
 overview 3:59–65
 standard-type integrable maps 3:61
 and Toda lattice 3:63
- Integrable dynamical systems, overview 3:107
- Integrable equations
 boundary-value problems for 1:346–351
 and non-Newtonian fluids 3:564
- Integrable evolution equations, boundary-value problems 1:350
- Integrable Hamiltonian system, and sine-Gordon equation 4:577
- Integrable many-body problems
 in arbitrarily dimensioned spaces 3:112
 overview 3:108
 and solitons 4:601
- Integrable maps, standard type 3:61
- Integrable nonlinear partial differential equations
 and inverse spectral transform technique 3:113
 properties 3:113
 and solitons 3:115
- Integrable partial differential equations
 and Bäcklund transformations 3:117
 classification techniques 3:120
 and conservation laws 3:118
 and nonlinear superposition principle 3:118
 overview 1:346–351
 properties 3:116
 and soliton ladder 3:118
- Integrable quantum field theory
 and affine Toda theories 3:55
 and conformal field theory 3:57
 minimal S -matrices 3:54
 overview 3:50–59
 and sine-Gordon theory 3:55
 and statistical systems 3:57
 types 3:58
- Integrable reductions, multidimensional quadrilateral lattices 3:83
- Integrable systems
 algebraic ansatz 2:427
 and algebraic geometry
 Hitchin systems 3:74
 isospectral deformations 3:68
 overview 3:65–78
 special functions 3:76
 symplectic reduction and r matrices 3:73
- Baker–Akiezer functions 2:431
- classical 3:50
- and classical differential geometry 3:78
- delta-bar approach
 data reductions 2:39
 Davey–Stewartson-II and Ishimori-I equations 2:38
 discrete systems 2:38
 explicit solutions 2:39
 KP-II hierarchy 2:36
 multidimensional problems 2:38
 nonlocal and local delta-bar problem 2:35
 nonsingular solutions 2:39
 Novikov–Veselov hierarchy 2:37
 overview 2:34–41
 quasiclassical limit 2:40
 Riemann surfaces 2:40
- and discrete geometry
 discrete surfaces 3:84
 overview 3:78–87
 quadrilateral lattice 3:79
- Dressing method 4:430
- and Frobenius manifolds 5:446
- functional equations
 overview 2:425–433
- geometric ansatz 2:430
- ground-state factorization 2:431
- and inverse scattering
 dromions 3:97
 lumps 3:97
 nonlinear Fourier transform 3:97
 overview 3:93–101
 solitons 3:96
 soliton theory development 3:95
 unification 3:100
- inverse-scattering transform 4:432
- Lax pairs 2:425
- Liouville integrable systems 4:534
- on manifolds 3:87–93
- nonlinear PDEs 3:113
- nonlinear steepest-descent method 4:433
- overview 3:106–123
- Poincaré algebra 2:428
- Poisson Lie group applications 1:517
- on Poisson–Nijenhuis manifolds 3:88
- quantum Calogero–Moser systems 2:428
- quantum completely integrable systems 2:148–155
- quantum Ruijsenaars–Schneider models 2:429
- in random matrix theory, overview 3:102–106
- Riemann–Hilbert methods
 overview 4:429–435
- topology 2:432
- and twistor theory 5:306
- Ward correspondence as 2:473
- q -Integral, in q -hypergeometric series 4:107
- Integral currents, and geometric measure theory 2:525
- Integral quantum Hall effect
 see also Von Klitzing effect
 discovery 2:403
- Integrals of motion, algebras, and recursion operators 3:91
- Integration-by-parts formulas, and loop space 3:415
- Interacting fields, and Euclidean field theory 2:261
- Interacting gases, Bose–Einstein condensates 1:314
- Interacting particle systems, and hydrodynamic equations
 entropy method 3:124
 hyperbolic equations 3:127
 nongradient models 3:126
 overview 3:123–129
 relative entropy method 3:128

- Interacting stochastic particle systems
 model systems 3:130
 nonreversible
 models 3:132, 3:134
 overview 3:130–135
 reversible models 3:131, 3:133
- Interaction energy, and abelian Higgs
 vortices 1:153
- Interaction-round-a-face model, and
 Yang–Baxter equations 5:467
- Interface-capturing methods, for
 multicomponent fluid flow 3:138
- Interface dynamics, interacting stochastic
 particle systems 3:132
- Interfaces
 and multicomponent fluids,
 overview 3:135–144
 statistical mechanics
 equilibrium crystals 5:61
 and Gibbs states 5:57
 overview 5:55–63
 pure phases and surface tension 5:56
 wetting phenomena 5:59
- Interface-tracking methods, for
 multicomponent fluid flow 3:136
- Interference, in continuous spontaneous
 localization model 4:272
- Intermittency
 and dynamical zeta function
 regularization 4:388
 and turbulence 3:144–151, 5:413
- Internal intermittency, in turbulence 3:144
- Interpolation methods
 for fluid mechanics 2:368
 mean-field spin glasses 3:410
- Intersection theory
 arithmetic intersection theory 3:155
 degeneracy loci application 3:154
 enumerative geometry application 3:154
 equivariant theories and stacks 3:156
 and Frobenius manifold 5:440
 moduli space application 3:155
 motives and higher Chow groups 3:155
 normal cone deformation 3:152
 overview 3:151–156
 refined intersections 3:152
 singularity theory application 3:154
 and topological sigma models 5:293
 topology links 3:153
 and 2D topological gravity 5:265
- Invariance
 N -point functions 1:236
 in quantum relative entropy 4:179
- Invasion percolation, characteristics 4:27
- Inverse mean curvature flow, general
 relativity 2:515
- Inverse problems
 Abel’s work 3:157
 and multidimensional Newton
 equation 3:158
 and multidimensional Schrödinger
 equation 3:159
 overview 3:156–160
 and wave propagation
 amplitude formula 1:344
 general setup 1:342
 model 1:342
 overview 1:340–345
 solutions 1:344
 wave coordinatization 1:343
- Inverse relations, and Yang–Baxter
 equations 5:472
- Inverse scattering
 dromions 3:97
 infinite Toda lattice 5:237
 and integrable systems
 nonlinear Fourier transform 3:97
 overview 3:93–101
 lumps 3:97
 for multidimensional Newton
 equation 3:158
 for multidimensional Schrödinger
 equation 3:159
 and solitons 4:595
 and soliton theory
 development 3:95
 unification 3:100
- Inverse-scattering transform method
 and integrable systems 4:432
 and nonlinear Schrödinger
 equations 3:557
- Inverse-spectral transformation
 for finite periodic Toda lattice 5:240
 for integrable PDEs 3:113
 and semi-classical trace formula 4:517
- Invertible cellular automata,
 overview 1:461
- Inviscid compressible fluid flows, Euler
 equations
 adiabatic equations 1:599
 entropy solutions 1:601
 isentropic equations 1:598
 local well-posedness 1:597
 multidimensional steady theory 1:599
 multidimensional unsteady
 problems 1:600
 overview 1:596
 singularity formation 1:597
- Inviscid flows
 incompressible perfect fluid 3:161
 overview 3:160–166
 perfect fluid equations 3:160
 and turbulence 3:165
- IQFT *see* Integrable quantum
 field theory
- IQHE *see* Integer quantum Hall effect
- IRF-vertex model, and Yang–Baxter
 equations 5:468
- Irreducible Green’s function, in
 perturbation theory 4:31
- Irreducible kernels
 and BS equation 4:484
 and Lippmann–Schwinger
 equation 4:483
 structure equations 4:484
- Isentropic Euler equations, inviscid
 compressible fluid flows 1:598
- Ishimori-I equations, and delta-bar
 problem method 2:38
- Ising models
 crystals 5:463
 equilibrium statistical mechanics
 overview 1:72
 in finite box 3:261
 holonomic quantum fields 2:660
 impure types 5:326
 Lenz–Ising model 5:328
 and phase transitions 4:55
 and QFT 5:326
- Isochronous systems
 deformed version 3:167
 examples 3:168
 overview 3:166–173
 transition to deterministic chaos 3:170
- Isomonodromic deformations
 and confluence 3:174
 experimental results 3:176
 overview 3:173–177
 and Schlesinger’s equation 3:173
- Isomorphisms, symmetric
 manifolds 1:503, 1:508
- Isoperimetric inequality, via proof with
 optimal transportation maps 3:633
- Isoperimetric Sobolev inequalities,
 properties 3:35
- Isospectral deformations, and integrable
 systems 3:68
- Isothermal sphere, in gravitational
 lensing 2:570
- Isotopies, in knot theory 3:401
- Isotropy groups, Lie groupoids 3:314
- Isotropy subgroups,
 and symmetry 5:184
- IST method *see* Inverse-scattering
 transform method
- Iterated multiscale approach,
 renormalization as 3:476
- Itô–Schrödinger equations, quantum
 dynamical semigroups 4:163
- ## J
- Jack polynomials, Macdonald
 polynomials 4:114
- Jackson’s q -Bessel function,
 equations 4:107
- Jackson’s sum, for terminating balanced
 ${}_8W_7$ 4:108
- Jackson’s ${}_8W_7$ summation formula, elliptic
 analog 4:115

- Jacobi diagrams
 - finite-type invariants of ZHS 2:350
 - Le–Murakami–Ohtsuki invariant 2:351
 - and Lie algebras 2:342
 - Jacobi fields, and hyperbolic billiards 2:717
 - Jacobi manifolds, integrable systems and recursion operators 3:87–93
 - Jacobi polynomials
 - equations 4:111
 - overview 3:643
 - Jacobi problem, separability 4:530
 - Jamiolkowski isomorphism, and quantum channels 1:474
 - Jang’s equation, and general relativity 2:506
 - Jones polynomial
 - and algebraic QFT 3:186
 - and braids 3:180
 - and conformal field theory 3:185
 - overview 3:179–187
 - statistical mechanics connections 3:184
 - and Temperley–Lieb algebra 3:181
 - and topological QFT 3:185
 - $V_L(t)$ original definition 3:183
 - Jones–Witten theory, overview 2:478
 - Jordan morphisms, between C^* -algebras 4:89
 - Jost–Lehmann–Dyson domain, and dispersion relations 2:99
 - Julia sets
 - in dynamics 2:400
 - and holomorphic dynamics 2:655
 - Jump discontinuities, symmetric–hyperbolic systems 5:163
- K**
- K3, definition 1:382
 - Kac–Moody Lie algebra, and solitons
 - affine Lie algebras 4:596, 4:599
 - Bäcklund transformation 4:596
 - ball–box systems 4:601
 - Drinfeld–Sokolov method 4:599
 - equation hierarchy 4:597
 - geometric interpretation 4:596
 - Hamiltonian formalism 4:597
 - Hirota method 4:597
 - and infinitesimal transformation groups 4:598
 - integrable many-body problems 4:601
 - inverse-scattering method 4:595
 - Kadomtsev–Petviashvili hierarchy 4:598
 - Krichever’s work 4:600
 - Lax representation 4:595
 - overview 4:594–602
 - quasiperiodic solutions 4:600
 - Sato’s work 4:598
 - similarity solutions 4:601
 - special solutions 4:600
- Kac–Moody Lie algebras
 - overview 1:676
 - and quivers 2:317–318
 - q -special function applications 4:116
 - Kac potentials, and phase transitions 4:57
 - Kadomtsev–Petviashvili hierarchy, and solitons 4:598
 - Kähler manifolds
 - characteristics 2:393
 - and Fedosov quantization 2:297
 - special holonomy 2:184
 - Kalikow’s condition, random walks in random environments 4:368
 - Kaluza–Klein models
 - brane worlds 1:369
 - cotangent bundle reduction 1:660
 - definition 1:382
 - KAM *see* Kolmogorov–Arnol’d–Moser theory
 - Kantorovich optimal transportation, basic problem 3:634
 - Kauffman bracket, and quantum gravity 3:218
 - Kawasaki dynamics, and interacting stochastic particle systems 3:131–132
 - Kawasaki–Glauber dynamics, and thermodynamic fluctuations 3:365
 - k -covectors, and geometric measure theory 2:524
 - KdV equation *see* Korteweg–de Vries equation
 - Kelvin’s theorem, incompressible perfect fluid 3:161
 - Kepler motion
 - classical gravitational many-body problems 2:576
 - classical mechanics overview 1:10
 - 2PI Kernel, and BS equation 4:484
 - 1PI Kernel, and Lippmann–Schwinger equation 4:483
 - Kernels
 - for generating function equation solution 1:559
 - and Lippmann–Schwinger equation 4:483
 - and microcausality 2:95
 - and QFT analytic program 4:469
 - Kerr metric, Einstein equation solutions 2:169
 - Kerr–Newman family, and black hole mechanics 1:301
 - Key rate, properties 2:235–236
 - K -group, C^* -algebras 1:395
 - Killing field, and black hole local equilibrium 1:302
 - Killing horizons theorem, and stationary black hole solutions 5:43
 - Killing vector, for stationary black holes 5:39, 5:41
- Kinematical Hilbert space, and 3D quantum gravity 4:647
 - Kinematic effects, Minkowski spacetime 1:100
 - Kinetic energy
 - and classical mechanics 1:2
 - equilibrium statistical mechanics overview 1:52
 - and peakons 4:17
 - Kinetic equations
 - averaging lemma 3:206
 - dispersive property 3:207
 - entropy and dissipation 3:207
 - entropy and irreversibility 3:201
 - linear billiard model 3:202
 - mathematical tools 3:206
 - and mean-field scaling 3:203
 - neutron transport equation 3:201
 - overview 3:200–208
 - from Schrödinger equation 3:205
 - semiconductor modeling 3:205
 - Kinetic models, Einstein equations
 - Boltzmann equation 2:199
 - distribution function and moments 2:199
 - Liouville–Vlasov equation 2:199
 - Kinetic theory
 - and multiscale approaches 3:476
 - and non-Newtonian fluids 3:564
 - Kingman theorem, for harmonic functions on Young graph 4:352
 - Kinks
 - and integrable PDEs 3:116
 - as solitons 4:603
 - Kirby moves, and Chern–Simons theory 4:499
 - Kirchhoff–Routh–Lin theory, and point vortex motion with boundaries 4:75
 - Kirwan injectivity, and Hamiltonian group actions 2:605
 - Kirwan surjectivity, and Hamiltonian group actions 2:605
 - KKS Poisson structure, and groupoids 3:315
 - Klein correspondence, and twistor theory 5:304
 - KMS condition *see* Kubo–Martin–Schwinger condition
 - Kneading determinants, and dynamical zeta function regularization 4:389
 - Kneading theory, and unimodal maps 5:344
 - Knot homologies
 - computations 3:210
 - construction 3:209
 - filtered complexes and deformations 3:209
 - geometric properties 3:210
 - H_2 construction 3:213
 - and Heegaard Floer homology 3:210

- overview 3:208–215
 - thin and thick knots 3:211
 - Knots
 - and Chern–Simons theory 4:499, 5:284
 - classical knot polynomials 2:341
 - and finite-type invariants 2:349, 2:350
 - in knot theory 3:221
 - and Kontsevich integral 3:235
 - and large- N dualities 3:273
 - and nonabelian BF theories 1:264
 - and nonabelian Chern–Simons theory 4:498
 - polynomial invariants 3:403
 - and quantum gravity, overview 3:215–220
 - sliced presentation, and Kontsevich integral 3:236
 - topological theory concepts 3:399
 - and vector fields 5:272
 - ZHS, finite-type invariants for 2:349
 - s -Knot states, and loop quantum gravity 3:340
 - Knotted tubes, magnetic flux 5:274
 - Knot theory
 - and braiding 3:227
 - invariants and applications 3:404
 - isotopies 3:401
 - knot and link polynomial invariants 3:403
 - knots, braids, bracket polynomial 3:221
 - loop transform and quantum gravity 3:227
 - notions and definitions 3:400
 - overview 3:220–231, 3:399–406
 - and quantum computing 3:227
 - Reidemeister moves 3:401
 - state summation 3:222
 - and topological QFT 3:227
 - topological role 3:400
 - Vassiliev invariants 3:404
 - Vassiliev invariants and rigid vertex graph invariants 3:223
 - Vassiliev invariants and Witten’s functional integral 3:225
 - Kobayashi–Hitchin conjecture, overview 2:473
 - Kodaira embedding theorem, and complex analysis 4:549
 - Kodaira–Spencer algebra, and manifold deformations 2:22
 - Koenigs lattice, properties 3:85
 - Kolmogorov–Arnol’d–Moser theory
 - lower dimensional tori 3:198
 - maximal tori 3:189
 - overview 3:189–199
 - physical applications 3:199
 - resummation scheme 1:30
 - stability 5:22
 - stability, and classical mechanics 1:20
 - and stability theory
 - Arnol’d diffusion and Nekhoroshev’s theorem 5:29
 - lower-dimensional tori 5:28
 - overview 5:26–32
 - PDE quasiperiodic solutions 5:28
 - stability *vs.* chaos 5:30
 - Kolmogorov–Arnol’d–Moser tori, quantization 3:584
 - Kolmogorov entropy, dissipative dynamical systems 2:108
 - Kolmogorov nondegeneracy, and Hamiltonian system stability 2:632
 - Kolmogorov theory
 - and Hamiltonian system permanent stability 2:633
 - and RPC3BP 3:191
 - for turbulence 5:300
 - Kondrakov theorem, and Sobolev space inequalities 3:34
 - Kontsevich conjecture, homological mirror symmetry 3:442
 - Kontsevich integral
 - arithmetic properties 3:236
 - chord diagrams, algebras 3:232
 - combinatorial construction 3:236
 - constituent explanation 3:233
 - for finite-type invariant fundamental theorem 2:343
 - for finite-type invariants 2:346
 - first terms 3:238
 - framed version 3:234
 - Hopf algebra structure 3:232
 - and knot operations 3:235
 - Kontsevich formula 3:233
 - Le–Murakami–Ohtsuki invariant 2:351
 - long chord omission 3:235
 - multiplicative properties 3:235
 - overview 3:231–239
 - universal Vassiliev invariant construction 3:234
 - weight systems 3:233
 - wheels formula 3:238
 - Korteweg–de Vries equation
 - approximation validity 3:241
 - bi-Hamiltonian formulation 1:292
 - boundary-value problems for 1:346
 - for breaking water waves 1:384–385
 - classical mechanics overview 1:14
 - derivation 3:240
 - and Fourier integral operators 3:43
 - integrability 3:243
 - mathematical aspects 3:244
 - and multi-Hamiltonian systems 3:461–462
 - overview 3:239–246
 - Koszul formula, and functional integration 2:439
 - Koszul–Tate differential, in BRST construction
 - with longitudinal complex 1:388
 - overview 1:388
 - Kowalevskaja’s gyroscope, classical mechanics overview 1:13
 - KP-II hierarchy, and delta-bar problem method 2:36
 - Kramers–Krönig relation, and linear response 2:91
 - Krawtchouk polynomials, characteristics 3:644
 - Kruskal–Szekeres spacetime, and stationary black holes 5:39
 - K -theory
 - Atiyah–Singer index theorem 3:248
 - C^* -algebras 3:249
 - Grothendieck–Riemann–Roch theorem 3:247
 - noncommutative tori 3:525
 - overview 1:552, 3:246–254
 - Thom class character form 3:391
 - twisted K -theory 3:251, 5:114
 - vector bundles 1:490
 - Kubo–Martin–Schwinger condition
 - and class field theory 3:603
 - and quantum central-limit theorems 4:139
 - and quantum spin systems 4:298
 - in thermal QFT 5:227
 - in Tomita–Takesaki modular theory 5:252
 - Kupka–Smale theorem, for C^∞ -generic dynamical systems 2:498
 - Kuramoto–Shivashinsky-perturbed Korteweg–de Vries equation, in fluid dynamics 1:284
 - k -vectors, and geometric measure theory 2:523
- ## L
- Laboratory frame, classical mechanics overview 1:12
 - Ladyženskaja–Babuška–Brezzi condition, in fluid mechanics 2:370
 - Lagrange inversion formula, for generating function equation solution 1:559
 - Lagrange nodes, in fluid mechanics 2:368
 - Lagrangian BRST construction, overview 1:390
 - Lagrangian dispersion
 - and multiparticle propagators 3:259
 - in nonsmooth flows 3:258
 - overview 3:255–260
 - passive scalar 3:259
 - single-particle diffusion 3:256
 - in smooth flows 3:256
 - Lagrangian formalism
 - BV quantization in 2:566

- Lagrangian formalism (*continued*)
 equations of motion 1:3
- Lagrangian manifold
 classical mechanics overview 1:8
 principal symbol on 5:48
- Lagrangian maps, and integrable discrete systems 3:60
- Lagrangian singularities, caustics as 3:623
- Lagrangian submanifolds
 and groupoids 3:315
 and optical caustics 3:623
- Lagrangian torus fibrations, and mirror symmetry 3:444
- Lagrangian variables
 in chiral perturbation theory 2:145
 in constrained systems 1:611–612
 and Hamiltonian fluid state 2:595
 in quantum chromodynamics 4:145
- Lag synchronization, chaotic systems 5:223
- Laguerre polynomials, overview 3:642
- Landau Fermi liquids,
 characteristics 2:304
- Landau–Ginzburg models, and derived categories 2:44
- Landau–Ginzburg–Wilson paradigm
 and quantum criticality 4:292
 and quantum phase transitions 4:294
 random matrix theory supersymmetry method 5:159
- Landau levels
 degeneracy 2:405
 filling factor 2:405
 lowest level problem 2:406
 overview 2:404
 Tsui–Stormer–Gossard effect 2:410
- Landau polynomial, and finite group symmetry breaking
 critical points 2:325
 overview 2:325
 reduction 2:326
- Landau singularity, and asymptotic causality 4:479
- Landscapes, stochastic resonance 5:89
- Laplacian
 and abelian BF theories 1:260
 capillary surface contribution 1:433
 multidimensional quadrilateral lattices 3:83
 on waves 1:341
- Large complex structure limit point
 and mirror symmetry setup 3:441
 monodromy around 3:447
 in two-torus 3:445
- Large deviation theory
 Ising model in finite box 3:261
 overview 3:261–263
- Large-eddy simulation, for turbulence theories 5:298
- Large Kähler limit point
 and mirror symmetry 3:441
 for two-torus 3:445
- Large- N duality
 background 3:270
 and Chern–Simons theory 3:270
 and closed-string theory 3:271
 history 3:270
 knots and links 3:273
 and open IIA prepotential 3:274
 and open invariants 3:273
 and open prepotential 3:273
 and open-string theory with branes 3:271
 overview 3:269–274
 and topological strings 3:266
 transitions 3:272
- Large- N expansion
 Chern–Simons theory 3:265
 and topological strings,
 overview 3:263–268
- Large quantum fields, in constructive QFT 1:627
- Larmor precession, classical mechanics overview 1:32
- Lattice gas automata, examples 1:465
- Lattice gauge theory
 anomalies and fermion doubling 1:211
 color confinement mechanisms 3:280
 finite temperature QCD 3:279
 mass spectrum 3:279
 numerical simulations 3:278
 overview 3:275–281
 quarks 3:277
 string tension 3:279
 topology 3:280
 Wilson’s formulation 3:276
- Lattice systems
 asymptotic lattices 3:85
 Calogero lattice 1:14
 cluster expansions 1:533
 in combinatorics 1:565
 equilibrium statistical mechanics overview 1:68
 gas automata examples 1:465
 Koenigs lattice 3:85
 and mean-field theory 3:412
 multidimensional integrable lattices 3:79
 multidimensional quadrilateral lattices 3:82
 and nilpotent Lie groups 4:96
 and percolative systems 4:25
 phase transition overview 4:376
 quantum lattice systems 4:131
 regularization, and perturbation theory 4:33
 root lattices 4:312, 4:313
 square lattices 4:311
 stochastic lattice gases 3:131–132
- Toda lattices 1:14, 3:63, 5:235–244
 and 2-step groups 4:98
- Law of large numbers, random walks in random environments 4:368
- Lax matrix
 in Calogero–Moser–Sutherland systems 1:405–406
 and integrable discrete systems 3:62
 Toda lattices 5:243
- Lax–Milgram theorem, and linear theory 2:220
- Lax pairs
 classical mechanics overview 1:14
 and integrable many-body problems 3:109
 for integrable pairs 2:425
- Lax representation, and solitons 4:595
- LBB condition *see*
 Ladyženskaja–Babuška–Brezzi condition
- LCLP *see* Large complex structure limit point
- LDT *see* Large deviation theory
- Lefschetz formula, for classical elliptic complexes 3:31
- Left-handed materials, for negative refraction 3:487
- Lehmann–Symanzik–Zimmermann theory
 reduction formulas in massive field theory 4:478
 and scattering theory 4:459
- Le–Murakami–Ohtsuki invariant
 definitions, diagrams, integrals 2:351
 recovery from 2:354
- Lenard–Magri relations, and bi-Hamiltonian formulation 1:292–294
- Length spectrum, in billiards 1:299
- Lens equation, and gravitational lensing 2:568
- Lensing, gravitational lensing 2:567–575
- Lens mapping, in gravitational lensing 2:570
- Lenz–Ising model, overview 5:328
- Lepton numbers, and symmetries 5:169
- Leptons
 annihilation 4:149
 anomalous magnetic moments 4:174
- Leray–Schauder theory
 and bifurcation theory 3:284
 and compact perturbations of identity 3:282
 and mapping degree,
 overview 3:281–285
- LERW *see* Loop-erased random walk
- LES *see* Large-eddy simulation
- Level-contour front-tracking methods, for multicomponent fluid flow 3:140
- Level-set methods, for multicomponent fluid flow 3:138, 3:140

- Levi-Civita, cyclic systems 4:373
- Levi theorem, and complex analysis 4:542
- Lévy transformations, multidimensional quadrilateral lattices 3:82
- L -functions, in arithmetic quantum chaos 1:218
- LGAs *see* Lattice gas automata
- LGW paradigm *see* Landau–Ginzburg–Wilson paradigm
- LHMs *see* Left-handed materials
- Liapunov–Schmidt reduction, in local bifurcation theory 1:277
- Lie algebras
 - characteristics 1:522
 - and compact Lie groups 1:577
 - decomposition 3:292
 - and deformation problem 2:21
 - and fermions and bosons 1:325
 - field embeddings 3:300
 - graded, definition 2:560
 - graphs 2:680
 - invariant subalgebra 3:290
 - and Jacobi diagrams 2:342
 - nonsemisimple algebra 3:291
 - p -forms 1:147
 - and Poisson groups 1:512
 - regular representation 3:291
 - semidirect sum algebra 3:290
 - semisimple, structure 3:293
 - semisimple algebra 3:291
 - simple algebra 3:291
 - and solitons 4:596, 4:599
 - solvable 3:290
 - structure 3:290
 - subfield restriction 3:300
 - for Toda lattices 5:240
 - as transformation groups 4:598
- Lie algebroids
 - cohomology 2:563
 - examples 3:317
 - integration 3:319
 - Lie groupoid 3:318
 - overview 2:562
 - Poisson groupoid 3:319
 - properties 3:318
 - symplectic groupoid 3:319
- Lie bialgebras
 - overview 1:511–518
 - special classes, and r -matrices 1:513
- Lieb–Thirring inequalities
 - characteristics 4:490
 - and b -pseudodifferential operators 2:703
- Lie groupoids
 - cotangent bundles 3:316
 - Lie algebroids 3:318
 - overview 3:312–320
 - properties 3:314
- Lie groups
 - block matrix decomposition 3:299
 - and canonical commutation relations 3:295
 - cohomology 1:551
 - Dynkin diagrams 3:296
 - linearization 3:287
 - matrix Lie algebras 3:289
 - matrix Lie groups 3:286
 - overview 3:286–304
 - and Poisson reduction 4:80
 - real forms overview 3:298
 - Riemannian symmetric spaces 3:301
 - root spaces 3:293
- Lie–Poisson brackets
 - and Hamiltonian fluid dynamics 2:594
 - and Hamilton’s equations 3:38
- Lie–Rinehart algebras
 - cohomology 2:563
 - overview 2:563
- Lie superalgebras
 - basic, representation theory 3:309
 - Cartan type 3:306
 - classical 3:306
 - overview 3:305–311
- Life span, metastable states 3:419
- Light
 - deflection in general relativity solar system tests 2:484
 - electromagnetism overview 1:50
- LIGO project, and gravitational waves 2:585
- Limit laws F_β , in random matrix theory 3:103
- Limit shape, Plancherel measure 4:348
- Limit theorems
 - F_β , in random matrix theory 3:103
 - in random matrix theory growth processes 2:590
- L–I model *see* Lenz–Ising model
- Lindstedt series
 - classical mechanics overview 1:18
 - and resonances 1:23
- Linear approximation, wavelets and polynomials 5:434
- Linear billiard model, kinetic equations 3:202
- Linear cocycles, and Lyapunov exponents 3:350
- Linear conditions
 - and QFT analytic program 4:468
 - in symmetric–hyperbolic systems 5:162
- Linear dispersion, absence in peakon singular solutions 4:14
- Linear equations, as partial differential equations 4:6
- Linear evolution equations
 - autonomous homogeneous equations 2:265
 - nonautonomous equations of hyperbolic type 2:269
 - nonautonomous equations of parabolic type 2:269
 - overview 2:265
- Linearized elasticity, elliptical differential equations 2:227
- Linear maps, and functional integration 2:437
- Linear momentum, and symplectic reduction 5:192
- Linear operators, spectral theory for
 - adjoint operator 4:634
 - compact operators 4:638
 - complexification 4:637
 - essential spectrum 4:639
 - hyponormal operators 4:641
 - normal operators 4:640
 - operational calculus 4:636
 - overview 4:633–645
 - and resolvent sets 4:635
 - spectral resolution 4:642
 - unbounded operators 4:639
- Linear parabolic equations, as partial differential equations 4:6
- Linear partial differential equations, and many-body problems 3:110
- Linear Penrose transform, zero-rest-mass free fields 5:313
- Linear programming, convex variational problems in duality 1:647
- Linear response theory
 - and dispersion relations 2:89
 - and quantum central-limit theorems 4:139
 - and relativistic quantum scattering 2:94
- Linear scalar field, and $*$ -algebras 4:203
- Linear stability analysis, and nonlinear Schrödinger equations 3:557
- Linear systems
 - periodic orbits 1:286
 - on Riemann sphere 4:436
 - on Riemann surfaces 4:422
- Linear theory, elliptic differential equations
 - biharmonic problem 2:222
 - Dirichlet problem for Poisson equation 2:221
 - general elliptic problems 2:223
 - Lax–Milgram theorem 2:220
 - linearized elasticity 2:227
 - method 2:217
 - mixed boundary conditions 2:222
 - motivation problem 2:216
 - Neumann problem for Poisson equation 2:221
 - Newton problem for general operators 2:222
 - overview 2:216–228
 - second-order elliptic problems 2:223
 - Sobolev spaces of positive order 2:217

- Linear theory, elliptic differential equations
(*continued*)
Stokes system 2:226
unbounded open 2:225
variational approach 2:220
- Line bundles
characteristics 1:492
and Fredholm operators 4:318
and Riemann surfaces 4:422
- Linked tubes, magnetic flux 5:274
- Linking numbers
in abelian BF theories 1:262
in knot theory 3:401
- Links
and Chern–Simons gauge theory 5:284
and Chern–Simons theory 4:499
homotopy links 2:347
and large- N dualities 3:273
magnetic links energy 5:276
and nonabelian Chern–Simons theory 4:498
polynomial invariants 3:403
quantum link invariants 2:353
topological theory concepts 3:399
- Liouville decomposition, and functional integration 2:437
- Liouville form, cotangent bundles 3:314
- Liouville integrable systems
and Calogero–Moser systems 4:123
separability 4:534
- Liouville–Vlasov equation, kinetic theory, Einstein equations 2:199
- Lippmann–Schwinger equation, and 1PI kernel 4:483
- Lipschitz case
solutions for microstructures 5:368
and stochastic differential equations 5:65
- Liquid crystals
and constitutive equations 3:565
dynamics 3:324
nematic phase elasticity 3:322
order parameter 3:322
overview 3:320–328
polymeric liquid crystals 3:565
smectic A phase elasticity 3:323
smectics defects 3:327
thermotropic and lyotropic systems 3:321
topological defects 3:324
uniaxial nematic phase defects 3:326
- Liquid–vapor transition, and van der Waals theory 4:56
- Little q -Jacobi polynomials, equations 4:111
- Ljusternik–Schnirelman theory
Banach space symmetric functionals 3:331
category 3:329
extensions 3:332
in infinite-dimensional manifolds 3:330
invariant functionals 3:331
minimax method 3:329
overview 3:328–333
- LKLP *see* Large Kahler limit point
- LLNs *see* Law of large numbers
- LMO invariant *see*
Le–Murakami–Ohtsuki invariant
- L nonconservation, and
Glashow–Weinberg–Salam theory 2:213
- Local bifurcation theory, overview 1:277
- Local boundary conditions, manifolds 3:29
- Local energy density, in general relativity 2:511
- Local equilibrium, in black hole mechanics 1:302
- Local existence, and incompressible Euler equations 3:12
- Locality, stochastic Loewner evolutions 5:82
- Localization, and equivariant cohomology 2:248
- Localization, for quasiperiodic potentials and Cantor spectrum 3:338
dynamical localization 3:337
Lyapunov exponents 3:334
multidimensional case 3:336
by nonperturbative methods 3:336
overview 3:333–338
perturbative and nonperturbative approaches 3:334
without Lyapunov exponents 3:336
- Local Lyapunov dimension, in dynamics 2:398
- Local statistics
in combinatorics 2:62
and thermodynamic limit 2:65
- Local symmetries
nonabelian BF theories 1:258
and quantum spin system symmetry 4:299
- Local triviality, singular reduction 5:196
- Local wavelet spectrum, and turbulence 5:410
- Local Weyl law, and quantum ergodicity and mixing 4:186
- Lockability, and entanglement measures 2:235
- Logarithmic function, example 3:639
- Logarithmic negativity, properties 2:235–236
- Logistic map, in dynamics 2:398
- London quantum hydrodynamics, and superfluids 5:117
- Longitudinal complex, in BRST construction
with Koszul–Tate differential 1:388
overview 1:388
- Long-range attraction, coalescence catastrophe from 1:85
- Long-time stability, quasiperiodic motions 1:26
- Loop algebras, for finite periodic Toda lattice 5:240
- Loop configurations, in combinatorics 1:570
- Loop-erased random walk, and stochastic Loewner evolutions 5:84
- Loop groups, as infinite-dimensional Lie group example 3:41
- Loop integrals
asymptotic expansion 4:37
Brown–Feynman reduction 4:38
master integrals 4:38
- Loop levels, definition 1:382
- Loop manifolds, and Feynman integral 4:516
- Loop quantum gravity
dynamics 3:342
and general relativity 3:339
operators and quanta of space 3:341
overview 3:339–343
quantum dynamics, overview 4:165–168
spin network and s -knot states 3:340
transform 3:227
- Loops, Le–Murakami–Ohtsuki invariant 2:351
- Loop space
heat measures on 3:414
integration-by-parts formulas 3:415
measures, overview 3:413–417
measures and absolute continuity 3:416
- Loop transform, and quantum gravity 3:227
- Loop variables, in nonabelian duality 2:205
- Lorentz distance, and nonspacelike cutlocus 3:344
- Lorentz forces
classical mechanics overview 1:32
electromagnetism overview 1:42
equation with differential forms 1:142
- Lorentz group, and Minkowski spacetime 1:96
- Lorentzian geometry
and geodesic incompleteness 3:346
Lorentz distance and nonspacelike cutlocus 3:344
overview 3:343–349
2D spacetimes 3:348
- Lorentzian groups, characteristics 4:102
- Lorentzian splitting theorem
definition 3:348
and geodesic incompleteness 3:346
- Lorentz invariance, and Einstein equivalence principle 2:483
- Lorentz surfaces, characteristics 3:348

- Lorenz attractor, in dynamics 2:399
 Lorenz-like attractors, and Lyapunov exponents 3:355
 Low-density limit, quantum dynamical semigroups 4:163
 Low-energy theories, and Seiberg dualities 4:504
 Lower-dimensional tori, and KAM 5:28
 Lower-semicontinuity
 in basic convex analysis 1:642
 in space BV 2:412
 Low-mass X-ray binaries, and gravitational waves 2:584–585
 Low temperature dynamics, and interacting stochastic particle systems 3:133–134
 LQG *see* Loop quantum gravity
 LSZ formalism *see*
 Lehmann–Symanzik–Zimmermann theory
 Lumps, characteristics 3:97
 Lusztig's quivers, characteristics 2:318
 Lyapunov exponents
 generic systems 3:354
 and Hénon-like attractors 3:355
 and linear cocycles 3:350
 and Lorenz-like attractors 3:355
 multiplicative ergodic theorem 3:350
 and nonuniform hyperbolicity 3:353
 and quasiperiodic potentials 3:334
 and random matrices 3:351
 and Schrödinger cocycles 3:352
 strange attractor characteristics 3:354
 strange attractor overview 3:349–356
 and subharmonic method 3:353
 Lyapunov stability
 fluid flow 5:4
 overview 3:350
- M**
 Macdonald–Koornwinder polynomials, overview 4:114
 Macdonald polynomials, for root systems 4:113
 MACHO *see* Massive Compact Halo Object
 Macroscopic dynamics, overview 3:358
 Macroscopic entropy, definition 1:382
 Macroscopic law, overview 2:374
 Macroscopic quantum systems, mathematical tools for 4:306
 Magnetic fields
 see also Electromagnetic fields
 divergence 1:44
 electron motion 4:246
 and fractional quantum Hall effect 2:409
 and Ginzburg–Landau energies 5:360
 and Ginzburg–Landau equations 2:549
 Magnetic flux, knotted and linked tubes 5:274
 Magnetic helicity, and knots 5:274
 Magnetic knots, energy 5:276
 Magnetic links, energy 5:276
 Magnetic moments, leptons 4:174
 Magnetic resonance imaging
 see also Nuclear magnetic resonance
 basic experiment 3:369
 and Bloch equation 3:367
 contrast and resolution 3:373
 overview 3:367–375
 selective excitation 3:370
 signal-to-noise ratio 3:372
 spin-warp imaging 3:371
 Magnetic Schrödinger operators, properties 4:490
 Magnetism
 see also Electromagnetism
 and negative refraction 3:485
 Magnetohydrodynamics
 basic modeling 3:375
 coupling 3:377
 electromagnetic description 3:376
 equations 3:378
 hydrodynamics description 3:376
 hyperbolic variant of equations 3:378
 industrial aluminum production 3:381
 numerical issues 3:380
 overview 3:375–383
 parabolic variant of equations 3:379
 simplified coupling 3:378
 Magnification, in gravitational lensing 2:569
 Majorana spinors, Clifford algebras 1:526
 Malkin's theorem, periodic orbits 1:287
 Malliavin calculus
 applications 3:387
 non-Euclidean setting variations 3:386
 overview 3:383–389
 quasi-sure analysis 3:385
 and stochastic differential equations 5:70
 Wiener space calculus 3:384
 Mandelbrot set is locally connected
 conjecture, and renormalization 5:347
 Mandelbrot sets
 and complex renormalization 5:346
 in dynamics 2:400
 and holomorphic dynamics 2:657
 Manifolds
 Calabi–Yau manifolds 1:399
 central manifolds 1:467–472, 3:473
 in classical mechanics 4:372
 classical Stein manifolds 1:502
 closed manifolds 2:177
 compact classical homogeneous manifolds 1:503
 compact complex manifolds 4:551–559
 compact manifolds 3:436
 compact Riemannian manifolds 3:36
 contact manifolds 1:631–636
 deformations, and Kodaira–Spencer algebra 2:22
 differential geometry overview 1:33
 and dispersion relations 2:97
 Einstein manifolds 1:180, 2:182–189
 and exceptional holonomy groups 4:444
 and Fedosov quantization 2:297
 and Feynman integral 4:516
 Feynman path integral on 4:9
 finite-type invariants 2:347
 flag manifolds 1:501
 formal deformation quantization 2:19
 four-manifold invariants 2:386–394, 2:457–463
 framed knots–links and Kirby moves 4:499
 Frobenius manifolds 5:438–447
 graded symplectic manifolds 2:564
 and Hamiltonian reduction of Einstein's equations 2:609
 and hyperbolic dynamical systems 2:723
 inertial manifolds 2:107
 infinite-dimensional manifolds 3:330
 Jacobi manifolds 3:87–93
 Lagrangian manifold 5:48
 local boundary conditions 3:29
 metric properties 1:35
 Poisson manifolds 1:511–518, 2:2, 3:37, 3:315, 4:79
 Poisson–Nijenhuis manifolds 3:88
 quantum three-manifolds 2:353, 4:117–122
 and Quillen determinant bundles 4:319
 Riemannian manifolds 1:550, 2:152, 4:527
 saddle-point manifold 5:158
 smooth manifolds 1:549, 1:550, 3:247
 spectral boundary conditions 3:30
 Stein manifolds 4:543
 stretching and squeezing 1:478
 supermanifolds 5:128–133
 symplectic manifolds 2:24–34, 3:38, 3:87–93, 3:314
 synchronization manifold 5:219
 three-manifold invariants 2:348–356, 2:477, 4:500
 toric manifolds 2:604
 Many-body systems
 classical gravitation 2:579
 integrable type 3:108
 and random matrix theory 4:344
 and solitons 4:601
 Map classes, rigidity property 3:436
 Mapping theory
 basic concepts 4:547
 maps in topologies 1:138
 Map T_L , rigorous renormalization group analysis 2:280

- Marginally trapped surfaces, and general relativity 2:505
- Marker-volume-of-fluid methods, for multicomponent fluid flow 3:140
- Markovian diffusion, renormalization analysis 3:479
- Markovian evolution, and microscopic and macroscopic dynamics 3:358
- Markov process, nonintersecting paths 2:51
- Markov random fields, for image processing 3:3
- Marsden–Weinstein reduction *see* Symplectic reduction
- Martens–Nowicki condition, and stochasticity 5:348
- Maslov’s index, and caustics 3:625
- Mass
- and Glashow–Weinberg–Salam theory 2:211
 - and Newtonian fluids 3:493
- Mass fraction, and Newtonian fluids 3:494
- Massive Compact Halo Object, as galactic microlens 2:572
- Massive field theory
- and analyticity 4:484
 - analytic N -point functions 4:478
 - and asymptotic causality 4:479, 4:484
 - and bound states 4:484
 - chronological functions 4:478
 - even theory 4:481
 - general analysis 4:481
 - higher spin in Minkowski spacetime 4:393
 - irreducible kernels 4:483
 - LSZ reduction formulas 4:478
 - models 4:482
 - nonlinear axiomatic program 4:480
 - overview 4:476
 - structure equations and asymptotic causality 4:486
 - two-body processes 4:480
 - two-particle bound states for even theories 4:485
 - wave equations for arbitrary spin 4:396
 - Wightman functions 4:478
- Massless fields
- higher spin in Minkowski spacetime 4:394
 - wave equations for arbitrary spin 4:397
- Massless particles, and scattering theory 4:462
- Mass scaling, and gravitational collapse 1:668, 1:670
- Mass spectrum, and lattice gauge theory 3:279
- Mass transport problems, duality 1:648
- Matched expansions, as multiscale approach 3:465–466
- Matched pair equations, and noncommutative spacetime 1:269
- Mathai–Quillen formalism
- and Atiyah–Jeffrey formula 3:396
 - Berezin integral 3:390
 - equivariant cohomology 3:392, 3:395
 - equivariant vector bundles 3:393
 - Euler class representatives 3:391
 - and gauged sigma model 3:395
 - overview 3:390–399
 - and superspace 3:394
 - and supersymmetric quantum mechanics 3:397
 - supertrace 3:390
 - Thom class in K -theory 3:391
 - Thom class representatives 3:391
 - and topological gauge theory 3:398
 - and topological sigma model 3:397
- Mathematical cosmology
- anisotropic models 1:656
 - approximate Friedmann–Lemaître models solutions 1:655
 - exact properties 1:653
 - Friedmann–Lemaître models 1:654
 - inhomogeneous models 1:656
 - origin of universe 1:657
 - overview 1:653–658
- Matrix algebras, and completely positive maps 4:92
- Matrix elements
- and quantum weak mixing 4:192
 - weak interaction 3:279
- Matrix families, as deformations 2:18
- Matrix Lie algebras, overview 3:289
- Matrix product states, characteristics 2:335
- Matrix theory, compactification 3:515
- Matrix types
- alternating-sign matrices 1:572
 - band matrices 5:243
 - block matrix decomposition 3:299
 - Cabibbo–Kobayashi–Maskawa matrix 5:36–37
 - classical r -matrices 1:511–518
 - commuting transfer matrices 2:158
 - corner transfer matrices 2:162
 - density matrix 1:82–83, 4:90
 - Lax matrix 1:405–406
 - products matrices 2:501, 4:332
 - r -matrices 1:185, 3:73
 - scattering matrix 4:257
 - Shur’s product of matrices 4:89
 - S -matrix 1:188, 3:54, 5:182, 5:466
 - transfer matrix method 1:564
- Matter, and Einstein equations
- charged fluids 2:199
 - dissipative fluids 2:198
 - dust 2:196
 - elastic media 2:200
 - fluid sources overview 2:196
 - kinetic models 2:199
 - overview 2:195–200
 - perfect fluid 2:196
 - and quantum group bundles 4:242
- Maupertuis’ principle, and classical mechanics 1:2
- Maurer–Cartan equation, and deformation problem 2:22
- Maximal Kolmogorov–Arnol’d–Moser theory tori
- and Arnol’d’s theorem 3:194
 - RPC3BP 3:191
- Maximal surfaces, and general relativity 2:507
- Maximal tori
- Cartan subalgebras 1:578
 - main properties 1:578
 - weights and roots 1:579
- Maximum principle, second-order elliptic problems 2:224
- Maxwell–Ampère equation, in magneto-hydrodynamics 3:376
- Maxwell–Coulomb equation, in magneto-hydrodynamics 3:376
- Maxwell equations
- and Abelian gauge group 3:40
 - differential forms 1:144
 - electromagnetism overview 1:46
 - and linear evolution equations 2:266–267
 - and magnetohydrodynamics 3:378–379
 - and symmetric–hyperbolic systems 5:164
- Maxwell–Faraday equation, in magneto-hydrodynamics 3:376
- Maxwell–Gauss equation, in magneto-hydrodynamics 3:376
- Maxwellian identities, adiabatic piston finite cylinder 1:168–169
- Maxwell’s rule, equilibrium statistical mechanics overview 1:63–64
- Mayer expansion, equations 1:531
- Mayer–Vietoris theorem, and singular homology 1:547
- Mean-field models
- ferromagnetic structure 4:657
 - glassy disordered systems 2:556
 - and multiscale methods 3:469
 - scaling and Vlasov equations 3:203
 - spin glasses 3:407
 - spontaneous symmetry breaking 4:140
 - and weakly coupled oscillators 5:452
- Mean-field spin glasses
- basic definitions 4:659
 - Ghirlanda–Guerra relations 3:412
 - Guerra’s integral representation 3:410
 - interpolation methods 3:410
 - lattice model applications 3:412
 - Neveu’s process 3:410

- overview 3:407–412
- random genealogies 3:410
- and short-range models 4:665
- site disordered models 3:408
- Mean-value inequality, and minimal submanifolds 3:423
- Measure rigidity, and arithmetic quantum chaos 1:217
- Mechanical systems
 - peakons as 4:15
 - simple, reconstruction phase symmetry 1:663
- Meissner solutions, Ginzburg–Landau equations 2:549
- Meixner–Pollaczek polynomials, characteristics 3:644
- Meixner polynomials, characteristics 3:644
- Melnikov integral
 - classical mechanics overview 1:26
 - geometric meaning 4:536
 - and separatrix splitting 4:536
- Melvin–Morton conjecture, for finite-type invariants 2:347
- Memory, and quantum channel capacity 1:430
- Memoryless quantum sources, and data compression 4:612
- Meshes, and fluid mechanics 2:368
- Meson scattering, dispersion relations 2:99
- Mesoscopic systems
 - Brownian motion 3:478
 - and random matrix theory 4:345
- Message-passing algorithms, from statistical physics 5:53
- Metastable free energies, and Pirogov–Sinai theory 4:62
- Metastable states
 - equilibrium states and dynamics 3:418
 - life span 3:419
 - overview 3:417–420
- Metastates, short-range spin glasses 4:571
- Metric-affine geometry, Einstein–Cartan theory 2:192
- Metric entropy, for random perturbations 4:335
- Metric independence, nonabelian BF theories 1:259
- Metricity, and Einstein–Cartan theory of gravitation 2:192
- Metric properties, manifolds 1:35
- Metric spaces, in topology 1:133
- Metric tensors, in general relativity 2:491
- MHD *see* Magnetohydrodynamics
- Microcausality
 - in axiomatic QFT 1:235
 - and kernels 2:95
 - N -point functions 1:236
 - and QFT analytic program 4:468
- Microcausality
 - galactic microlensing 2:571
 - quasar microlensing 2:573
- Microlocal analyticity properties, and QFT analytic program 4:471
- Micro–macro models, non-Newtonian fluids 3:564
- Microscopic dynamics, overview 3:358
- Microscopic entropy
 - and D-branes and black branes 1:377
 - definition 1:382
 - extremal three-charge system 1:378
- Microscopic theory, Brownian motion 3:478
- Microstructures, variational techniques for
 - Austenite–Martensite transformations 5:363
 - calculus of variations 5:364
 - minimizers 5:366
 - overview 5:363–368
 - relaxation 5:365
- Mimimal supersymmetric standard model
 - characteristics 5:141–142
 - Higgs sector 5:142
- Minimal-area problem, for open string field theory 5:99
- Minimal graphs, and minimal submanifolds 3:421
- Minimally coupled scalar, definition 1:382
- Minimal S -matrices, and integrable QFT 3:54
- Minimal standard model, electroweak theory 5:34–35
- Minimal submanifolds
 - area-minimizing surfaces 3:428
 - Bernstein and Bers’ theorem 3:424
 - classical results 3:420
 - embedded minimal disks 3:425
 - Gauss map 3:421
 - Heinz’s curvature estimate 3:425
 - maximum principle 3:422
 - minimal graphs 3:421
 - minimal surface applications 3:429
 - minimal surface construction 3:427
 - minimal surface min–max construction 3:428
 - monotonicity and mean-value inequality 3:423
 - overview 3:420–432
 - plateau problem 3:427
 - and Rado’s theorem 3:424
 - regularity theory 3:426
 - and Simons inequality 3:425
 - stable minimal surfaces 3:426
 - Weierstrass representation 3:427
- Minimal surfaces
 - applications 3:420
 - in basic convex analysis 1:644
 - and black holes 3:430
 - and constant mean curvature surfaces 3:430
- construction 3:427
- and general relativity 2:503, 2:504
- min–max construction 3:428
- and positive-mass theorem 3:429
- Ricci flow finite extinction 3:431
- Minimax principle
 - in calculus of variations, overview 3:432–439
 - and closed geodesics 3:436
 - compact energy hypersurfaces 3:437
 - general principle 3:435
 - map class rigidity 3:436
 - mountain-pass theorem 3:432
 - and nonlinear elliptic boundary-value problem 3:433
 - and saddle point problems 4:448
- Minimizers, and microstructures 5:366
- Minimum dissipation principle, and Hamilton–Jacobi equation 3:363
- Minisuperspace approximation, in quantum cosmology 4:154
- Minkowski spacetime
 - and charged particles 1:105
 - conformal infinity 1:222
 - and Einstein equations solutions 2:166
 - and electromagnetic fields 1:105
 - geometry, particles, dispersion relations 2:94
 - higher spin 4:393
 - kinematic effects 1:100
 - and Lorentz group 1:96
 - low spin wave equations 4:392
 - and 4-momentum 1:102
 - overview 1:96–109
 - and particles 1:102
 - physical interpretation 1:97
 - and relativity principle 1:101
 - and spacetime boundaries 1:326
 - stability data formulation 5:15
 - stability overview 5:14–19
 - stability problems 5:19
 - stability variations 5:16
 - strong peeling 5:18
- Min–max construction, minimal surfaces 3:428
- Mirror symmetry
 - and A and B topological sigma models 5:295
 - homological 3:442
 - Lagrangian torus fibrations 3:444
 - monodromy around LCLP 3:447
 - quintic 3-fold 3:446
 - Strominger–Yau–Zaslow conjecture 3:443
 - and SYZ conjecture 1:401
 - and two-torus 3:444
- Miura transformations, and Painlevé equations 4:3

- Mixing
 - and ergodic theory 2:253
 - and vortex dynamics 5:394
- MK optimal transportion *see* Monge–Kantorovich optimal transportation problem
- MLC *see* Mandelbrot set is locally connected conjecture
- Mode interactions
 - and Birkhoff normal form 5:189
 - and Hopf bifurcation 5:189
- Modified Bessel function, characteristics 3:639
- Modular automorphism group, Tomita–Takesaki modular theory
 - algebra and states 5:252
 - KMS condition 5:252
 - overview 5:252
- Modular categories, characteristics 1:355
- Modular functor, and TQFT 1:359
- Modular invariance, in conformal field theory 1:335, 5:321
- Modular nuclearity, and Tomita–Takesaki modular theory 5:255
- Modular position, and Tomita–Takesaki modular theory 5:256
- Modular tensor categories, overview 1:351–359
- Modular theory, and von Neumann algebras 5:383
- Modulation equations
 - in fluid dynamics 1:284
 - non-Korteweg–de Vries equation 3:245
- Modules
 - in topologies 1:139
 - and twisted K -theory 3:252–253
- Moduli space
 - and A–B flat connections 1:196
 - cohomology 1:551
 - Einstein manifolds 2:184
 - fine and coarse 3:452
 - and gauge theory 2:470
 - as intersection theory application 3:155
 - as orbit spaces 3:454
 - overview 3:449–458
 - problem specifications 3:450
 - stabilization, and string theory 5:109
 - topology 2:474
 - vector bundles 3:456
- Molecules, equilibrium statistical mechanics overview 1:51
- Møller operators, and time-dependent scattering theory 4:252
- Moment maps
 - basic properties 2:601
 - in gauge theory 2:473
 - and symplectic reduction 3:73
- Moment polytope, and Hamiltonian group actions 2:604
- Moments
 - adiabatic piston 1:165
 - kinetic theory, Einstein equations 2:199
- 4-Momentum, and Minkowski spacetime 1:102
- Momentum, in fluid mechanics 2:365
- Momentum cutoff regularization, perturbation theory 4:32
- Momentum frame, classical mechanics overview 1:12
- Momentum maps
 - peakon reduction as 4:19
 - properties 5:192
 - and symplectic reduction 5:192
- Momentum transfer variables, and analyticity 4:472
- Monge–Kantorovich optimal transportation problem, basic problem 3:634–635
- Monge optimal transportation, basic problem 3:634
- Monochromatic plane waves, electromagnetism overview 1:48
- Monopoles
 - characteristics 4:606
 - electric–magnetic duality 2:202
 - and noncommutative instanton 3:522
 - and Seiberg–Witten invariants 2:461
- Monotonicity
 - and entanglement measures 2:235
 - and Jacobi fields 2:717
 - and minimal submanifolds 3:423
 - and wave fronts 2:718
- Montel’s theorem, and complex analysis 4:546
- Moreau–Fenchel conjugate, in basic convex analysis 1:643
- Morita equivalence
 - in noncommutative Yang–Mills theory 3:527
 - star products in deformation quantization 2:14
- Morita invariance, C^* -algebras 3:250
- Mori–Zwanzig projection techniques, overview 3:473
- Morphisms, between C^* -algebras 4:89
- Morse theory
 - and Floer homology 2:356
 - reformulation for study 2:359
- Motivation
 - and linear theory 2:216
 - and stochastic differential equations 5:63
- Motives, and intersection theory 3:155
- Motivic Galois theory, and renormalization 3:606
- Motzkin path, in combinatorics 1:555
- Mountain-pass theorem
 - and nonlinear elliptic boundary-value problem 3:433
 - overview 3:432
- MRFs *see* Markov random fields
- MRI *see* Magnetic resonance imaging
- MSM *see* Minimal standard model
- M spacetimes *see* Minkowski spacetime
- MSSM *see* Minimal supersymmetric standard model
- M-theory
 - and dualities 5:137
 - overview 5:108
 - in quantum corrections to gauge theory construction 1:365
- Multicomponent fluid flow
 - future work 3:143
 - hybrid methods 3:140
 - interface-capturing methods 3:138
 - interfaces, overview 3:135–144
 - interface-tracking methods 3:136
 - phase-field method 3:140
- Multidimensional cellular automata, overview 1:459
- Multidimensional CWT, wavelets applications 5:422
- Multidimensional integrable lattices
 - analytic methods 3:80
 - discrete geometry 3:79
- Multidimensional Newton equation, inverse scattering for 3:158
- Multidimensional nuclear magnetic resonance, and evolution domain and RF pulses 3:597
- Multidimensional quadratures, classical mechanics overview 1:9
- Multidimensional quadrilateral lattices
 - Darboux-type transformations 3:82
 - distinguished integrable reductions 3:83
- Multidimensional Schrödinger equation, inverse scattering for 3:159
- Multidimensional steady theory, inviscid compressible fluid flows 1:599
- Multidimensional unsteady problems, inviscid compressible fluid flows 1:600
- Multifractal analysis
 - dynamical system functions and measures 2:397
 - wavelets applications 5:421
- Multigrid algorithms, and multiscale approaches 3:474
- Multi-Hamiltonian systems
 - examples 3:460
 - overview 3:459–465
- Multiparticle propagators, and Lagrangian dispersion 3:259
- Multipartite stars, and entanglement 2:232

- Multiple radiofrequency pulses, and multidimensional NMR 3:597
- Multiplicative ergodic theorem and Lyapunov exponents 3:350 and random dynamical systems 4:332
- Multiplicative models in topological QFT 1:232 for turbulence intermittency 3:148
- Multiplicative subtraction, and Hopf algebra structures 2:682
- Multiplicity in optical caustics 3:626 tensor product, and compact groups 1:585
- Multiply-imaged quasars, characteristics 2:572
- Multiresolution analysis discrete wavelets 5:433 orthogonal wavelets 5:429
- Multiscale approach abstract examples 3:467 anomalous diffusion 3:480 boundary layers and matched expansions 3:474 Brownian motion 3:478 classical example 3:467 in constructive QFT 1:629–630 diffusion case 3:477 domains of application 3:468 effective diffusion 3:480 and Markovian diffusion 3:479 mean-field approaches 3:469 overview 3:465–482 in perturbation theory 2:58 and renormalization group 3:476 scale invariance 3:470 scales and regimes 3:469 scaling limits 3:470 self-avoiding walk renormalization analysis 3:480 singular perturbations 3:470 singular perturbations and secular divergences 3:466 and slow-fast variables aggregation techniques and coarse-grainings 3:473 amplitude equations 3:471 averaging techniques 3:472 central manifold theorem 3:473 decomposition 3:471 numerical aspects 3:474 parametric approximation 3:471 projection techniques 3:473 quasistationary approximation 3:472 slow invariant manifolds 3:473 stochastic approaches 3:471 technical principles 3:466 wavelets applications 5:422 and weakly correlated diffusion 3:479
- Multivector representations, Clifford algebras 1:527
- Muonium, quantum electrodynamics 4:173
- Murray–von Neumann classification, von Neumann algebras 5:380
- ## N
- Nakajima’s quivers, characteristics 2:320
- Nambu space, and disordered superconductors 5:209
- Natural boundary conditions in fluid mechanics 2:365 and 2D Ising model 5:326
- Naturality in K-theory 3:247 in topological QFT 1:232 and twisted K-theory 3:253
- Natural transformations, and algebraic QFTs 1:203
- Navier–Stokes equations fluid flow stability 5:3 in fluid mechanics 2:373 for stochastic hydrodynamics 5:75 viscous compressible fluid flows, overview 1:602
- N-body Schrödinger operators, properties 4:491
- NC sector *see* Neutral-current sector
- Near-extremal entropy, and branes 1:379
- Nearly integrable dynamical systems, hyperbolic system comparison 2:126
- Negative refraction approaches 3:491 and artificial magnetism 3:485 and artificial plasma 3:484 index media 3:487 overview 3:483–491 perfect lens 3:488
- Nekhoroshev theorem and Arnol’d diffusion 5:29 and Hamiltonian systems 2:632 overview 5:23 stability theory 4:592 steepness, and Hamiltonian system stability 2:632
- Nematic droplets, topological defects 3:325
- Nematic liquid crystals, and topological defects 5:262
- Nematic phase, liquid crystals, elasticity 3:322
- Neumann problem, for Poisson equation 2:221
- Neural networks, overview 3:407–412
- Neutral currents, and Glashow–Weinberg–Salam theory 2:211
- Neutral-current sector, electroweak theory 5:35
- Neutral pion decay, chiral anomaly application 1:207
- Neutron stars, and gravitational waves 2:584
- Neutron transport equation, diffusion limit 3:201
- Neveu–Schwarz states, definition 1:382
- Neveu’s process, mean-field spin glasses 3:410
- Newhouse phenomenon, and homoclinic bifurcation 4:592
- Newton equation, inverse scattering for 3:158
- Newton flows, and semi-classical trace formula 4:513
- Newtonian cosmology, overview 3:507
- Newtonian dynamics, relativity 1:41
- Newtonian fluids, thermohydraulics adjoint system 3:497 Boussinesq approximations 3:493 Boussinesq local balances 3:493 density state equations 3:493 dimensionless local balances 3:495 direct system stability 3:496 governing equations 3:492 gravitational buoyancy convection 3:498 hydrostatic state 3:494 hydrostatic state stability 3:497 linear stability 3:496 nonlinear stability 3:497 overview 3:492–502 thermocapillary convection 3:502
- Newtonian potential, classical mechanics overview 1:10
- Newton’s theory of gravitation Cartan–Friedrichs formalism 3:504 in general relativity solutions 3:507 global laws 3:506 local laws 3:505 overview 3:503–509
- NGT *see* Newton’s theory of gravitation
- Nielsen–Olesen vortex string, quantum fields with topological defects 4:228
- Nilpotent Lie groups inner product and signature 4:95 lattices and completeness 4:96 overview 4:94 pseudo-Riemannian nilpotent Lie groups 4:94–103 2-step groups 4:95
- NLS equations *see* Nonlinear Schrödinger equations
- Noether theorem, and classical mechanics 1:1–2
- No-hair theorem definition 1:382 stationary black holes 5:43

- No invariant line fields conjecture, and holomorphic dynamics 2:657
- Noise
- abstract framework 4:332
 - and coherent structures from turbulent flows 5:414
 - in dynamics 4:330
 - random maps 4:331
 - random perturbations of flows 4:331
- Nonabelian BF theories
- and Chern–Simons theory 1:263
 - classical action 1:258
 - classical moduli space 1:258
 - classical nonabelian BF theories 1:262
 - deformed 1:263
 - diffeomorphisms 1:258
 - equations of motion 1:258
 - gauge invariance 1:258
 - and gravity 1:263, 1:264
 - knots and generalized invariants 1:264
 - local p -form symmetries 1:258
 - local symmetries 1:258
 - metric independence 1:259
- Nonabelian Chern–Simons theory
- knots, links, braids 4:498
 - link invariants 4:498
 - overview 4:496
- Nonabelian chiral anomalies, characteristics 1:205
- Nonabelian duality
- treatments 2:208
 - via loop variables 2:205
- Nonabelian effects, A–B effects 1:198
- Nonabelian gauge theory
- with differential forms 1:141–150, 1:146
 - Dirac fields, overview 2:67–74
 - as infinite-dimensional Lie group
 - example 3:41
- Nonamenable graphs, percolation on 4:26
- Nonautonomous equations
- hyperbolic type 2:269
 - parabolic type 2:269
- Nonaxial maximal isotropy subgroups, and equivariant bifurcation theory 5:187
- Nonbipartite graphs, in combinatorics 2:63
- Noncanonical Hamiltonian structure
- Lie–Poisson brackets 2:594
 - overview 2:593
 - Poisson brackets 2:593
- Noncanonical Poisson brackets
- from fluid Hamiltonian 2:598
 - and Hamiltonian fluid
 - dynamics 2:593
 - and ideal Hamiltonian fluid 2:599
- Noncommutative Fourier transform, on compact group 1:583
- Noncommutative gauge theory
- open Wilson line 3:519
 - overview 3:519
- perturbative dynamics 3:521
 - Seiberg–Witten map 3:520
- Noncommutative geometry
- associated topics 3:523
 - Bismut–Chern character 4:10
 - characteristics 3:511
 - and D-branes 3:522
 - Fermionic Brownian motion 4:11
 - Feynman path integral on 4:9
 - GMS solitons 3:521
 - infinite-dimensional volume
 - element 4:9 - and matrix theory
 - compactification 3:515
 - open-string in B -field 3:515
 - overview 3:509–515
 - path-integral overview 4:8–12
 - and quantum groups 2:699
 - and quantum Hall effect 4:247
 - quantum mechanics 3:512
 - spectral action 3:513
 - spectral triples 3:512
 - from string theory, overview 3:515–524
- Noncommutative instantons, and monopoles 3:522
- Noncommutative quantum field theory
- IR–UV mixing 3:518
 - overview 3:517
 - planar and nonplanar diagrams 3:518
- Noncommutative random variable, freeness 2:417
- Noncommutative spacetime
- and bicrossproduct Poincaré quantum groups 1:272
 - and cogravity 1:265
 - matched pair equations 1:269
 - overview 1:265–274
 - and Planck-scale Hopf algebra 1:270
 - and Poisson–Lie T -duality 1:273
 - and standard model 2:141
- Noncommutative tori
- algebra of functions 3:524
 - connections 3:525
 - and instantons 3:529
 - and Morita equivalence 3:527
 - overview 3:524–530
 - projective modules 3:524
 - K -theory 3:525
 - Yang–Mills theory on, overview 3:526
- Noncompact Riemannian symmetric manifolds
- overview 1:504
 - R -spaces 1:508
 - symmetric cones 1:505
- Nonconserved systems, phase ordering
- domain growth 4:48
 - interface dynamics 4:49
 - overview 4:47
- Nonconvex functional, on bounded variation 2:415
- Noncrossing partitions, in free probability theory 2:420
- Nondegeneracy, and semi-classical trace formula 4:514
- Nondeterministic polynomial time, and combinatorial problems 5:50
- Nonequilibrium fluctuation dissipation relation, derivation 3:361
- Nonequilibrium statistical mechanics
- chaotic hypothesis 3:533
 - dynamical systems approach 3:540
 - ensembles and equivalence 3:537
 - fluctuations 3:536
 - fractal attractors, pairing, time reversal 3:537
 - future work 3:538
 - heat, temperature, entropy
 - production 3:534 - numerical simulations 3:546
 - overview 3:530–540
 - theory–numerical simulations, overview 3:544–551
 - thermodynamic fluxes and forces 3:535
 - thermodynamics 3:532
- Nonequilibrium steady state, and interacting stochastic particle systems 3:134
- Nonergodicity, and lattice system phase transitions 4:377
- Non-even theories, and two-particle bound states 4:486
- Nonextremal black holes, and branes 1:379
- Nongradient models, interacting particle systems 3:126
- Nonintersecting paths, in combinatorics 1:569
- Nonlinear approximation, wavelets 5:434
- Nonlinear axiomatic program, and massive field theory 4:480
- Nonlinear dynamical semigroups, overview 4:164
- Nonlinear elliptic boundary-value problem, and mountain-pass theorem 3:433
- Nonlinear equations, twistor theory
- anti-self-dual Yang–Mills equations 5:306
 - graviton construction 5:305
 - overview 5:305
- Nonlinear evolution equations
- for fluid flow stability 5:3
 - fully nonlinear, parabolic type 2:271
 - hyperbolic quasilinear evolution
 - equations 2:270 - overview 2:270
 - parabolic quasilinear evolution
 - equations 2:271
- Nonlinear Fourier transform
- direct problem 3:98
 - inverse problem 3:99
 - and inverse scattering method 3:97
 - scattering data time dependence 3:99

- Nonlinear harmonic oscillators,
characteristics 3:112
- Nonlinear instability, fluid flow 5:4
- Nonlinear model equations, for breaking
water waves 1:384
- Nonlinear optics, and Schrödinger
equations 3:552, 3:555
- Nonlinear partial differential equations,
integrability 3:121
- Nonlinear Penrose transform, curved
twistor spaces 5:314
- Nonlinear problems, symmetric-hyperbolic
systems 5:163
- Nonlinear Schrödinger equations
boundary-value problem 1:349
discrete NLS systems 3:559
and Ginzburg-Landau equations 3:552
inverse-scattering transform 3:557
and linear stability analysis 3:557
in nonlinear optics 3:555
in optical communications 3:556
and optics 3:552
overview 3:552-560
universal character 3:553
vector NLS systems 3:558
in water waves 3:554
and wave collapse 3:558
- Nonlinear sigma models
random matrix theory supersymmetry
method 5:159
and renormalization groups 4:414
- Nonlinear steepest-descent method, and
integrable systems 4:433
- Nonlinear superposition principle, and
integrable PDEs 3:118
- Nonlocal signaling
and deterministic hidden-variables
theories 4:269
and quantum nonequilibrium 4:268
- Non-Newtonian fluids
behavior types 3:561
constitutive equations 3:563
continuum mechanics 3:562
modeling overview 3:562
objectivity and frame invariance 3:563
overview 3:560-567
viscoelastic flows 3:566
- Nonoptical caustics,
characteristics 3:627
- Nonorthogonal states, in quantum
theory 4:269
- Nonperturbative renormalization
in constructive QFT 1:626
for fermions 4:413
- Nonrelativistic current algebra,
overview 1:677
- Nonrelativistic quantum chromodynamics,
overview 2:144
- Non-renormalizability,
characteristics 4:404
- Non-Seifert manifolds, and Hamiltonian
reduction of Einstein's
equations 2:609
- Non-semimartingales, and stochastic
differential equations 5:70
- Nonsemisimple Lie algebras,
characteristics 3:291
- Nonsingular isothermal sphere, in
gravitational lensing 2:570
- Nonsmooth flows, two-particle Lagrangian
dispersion 3:258
- Nonspacelike cutlocus, and Lorentz
distance 3:344
- Nontrivial scattering, and QFT with
indefinite metric 4:219
- Nonuniform hyperbolicity
and Lyapunov exponents 3:353
overview 2:728
- Nonuniqueness, capillary
surfaces 1:436
- Nordtved effect, in general relativity solar
system tests 2:484
- Normal cone, algebraic intersection
products 3:152
- Normal currents, and geometric measure
theory 2:525
- Normal fluctuations, quantum lattice
systems 4:133
- Normal forms
Birkhoff normal forms 2:150, 5:189
and Hamiltonian group actions 2:603
and Hamiltonian system stability 2:632
resonant case 3:583
and semiclassical approximation,
overview 3:578-585
singularity and bifurcations 4:591
- Normalization
in K -theory 3:247
in topological QFT 1:232
and twisted K -theory 3:252-253,
3:253
- Normal operators, and spectral
theory 4:640
- Normal stresses, and non-Newtonian
behavior 3:561
- Normed space, compact perturbations of
identity 3:282
- No-slip boundary conditions, in fluid
mechanics 2:365
- Novikov-Veselov hierarchy, and delta-bar
problem method 2:37
- No wandering domains theorem,
holomorphic dynamics 2:655
- NP *see* Nondeterministic polynomial time
- N-particle quantum scattering
long-range interactions 3:592
nonperturbative approach 3:590
overview 3:585-592
problem setting 3:586
resolvent equations for 3:588
- N-point functions
and axiomatic QFT 1:236
in massive field theory 4:478
and QFT analytic program 4:468
in thermal QFT 5:231
- N point vortices, velocity field tracers 4:73
- NRQCD *see* Nonrelativistic quantum
chromodynamics
- NS branes, in string theory 1:376
- NS5 branes, and gauge theories 1:363
- NSEs *see* Navier-Stokes equations
- Nuclear C^* -algebras, and injective von
Neumann algebras 4:93
- Nuclear magnetic resonance
see also Magnetic resonance imaging
and evolution domain and RF
pulses 3:597
excitation and detection 3:593
overview 3:592-600
solid-state NMR 3:599
spin ensemble statistics 3:593
spin Hamiltonian 3:596
spin relaxation 3:595
- Nuclear physics, and effective field
theory 2:146
- Nuclear spin Hamiltonian,
characteristics 3:596
- Nucleation
droplet free energy and experiments 4:51
overview 4:50
rate, and droplet size distribution 4:51
- Null hypersurfaces, and general
relativity 2:508
- Null Raychaudhuri equation, and
spacetime topology 4:620
- Number theory
class field theory 3:603
dilogarithm, polylogarithms, zeta
values 3:601
Feynman graphs 3:602
Galois theory 3:602
motives 3:605
overview 3:600-607
- Numerical methods
classical gravitational many-body
problems 2:579, 2:581
for multicomponent fluid flow 3:136
in vortex dynamics 5:396
and wavelets applications 5:425
- Numerical simulations
adiabatic piston 1:169
lattice gauge theory 3:278
and nonequilibrium statistical mechanics,
overview 3:544-551
- O**
- Oblique wave front collisions,
peakons 4:19
- Observability, as control problem 1:636

- Ocean primitive equations
 boundary conditions 2:536
 overview 2:535
 strong solutions 2:537
- Odd dimensions, Clifford algebra
 representations 1:525, 1:526
- ODEs *see* Ordinary differential equations
- ODLRO *see* Off-diagonal long-range order
- Off-diagonal long-range order
 and quantum spin system phase transitions 4:300
 quantum statistics overview 1:83
- Off-equilibrium dynamics, and glassy disordered systems 2:557
- Off-shell states, and operads 3:613
- Oldroyd models, and non-Newtonian fluids 3:564
- One-parameter groups, functional analysis
 overview 1:93
- 1PI *see* One-particle irreducible
- One-particle Dirac theory, overview 2:80
- One-particle irreducible, in perturbation theory 4:31
- One-particular systems, Euclidean field theory 2:257
- One-point functions
 eight-vertex model 2:162
 hard-hexagon model 2:163
- One-way distillable entanglement, properties 2:235–236
- On-shell states, and operads 3:613
- OP *see* Order parameter
- OPE *see* Operator product expansion
- Open bosonic string field theory
 classical action 5:96
 closed strings 5:100
 and minimal area metrics 5:99
 open-string field 5:95
 overview 5:95
 quantization 5:98
- Open-closed homotopy algebras, operads
 in 3:612
- Open-closed string field theory,
 overview 5:101
- Open four-manifolds, exotic structures
 on 2:392
- Open prepotential
 and large- N dualities 3:273, 3:274
 overview 3:273
- Open-string B model, sheaves and derived categories 2:41
- Open-string theory
 B -field 3:515
 with branes 3:271
 classical solutions 5:101
 Seiberg–Witten limit 3:517
 vacuum SFT 5:102
- Open Wilson line, in noncommutative gauge theory 3:519
- Operads
 in algebra 3:611
 in A_∞ -algebras 3:611
 in L_∞ -algebras 3:611
 and Batalin–Fradkin–Vilkovisky construction 3:614
 and Batalin–Vilkovisky construction 3:614
 in BV algebras 3:615
 A_∞ -categories 3:615
 definitions and examples 3:609
 and deformation quantization 3:614
 in Gerstenhaber algebras 3:613
 in homotopy theory 3:610
 insertion operads 3:615
 in mathematical physics 3:613
 on-shell *vs.* off-shell states 3:613
 in open-closed homotopy algebras 3:612
 overview 3:609–616
 in Poisson algebra 3:612–613
 and PROPs 3:615
- Operational calculus, and spectra theory 4:636
- Operator algebras
 functional analysis overview 1:95
 vertex operator algebras 5:317–322
- Operator product expansion, in QFT
 conformal field theory 3:619
 free field theory 3:617
 interacting theories 3:617
 overview 3:616–620
 perturbative example 3:618
 two dimensions 3:620
 Ward identities 3:619
- Operator spaces, topologies 1:135
- Operator time, and quantum ergodicity 4:190
- Operator types
 adjoint operator 4:634
 almost-Mathieu operator 3:335
 codifferential operator 1:142
 compact operators 4:638
 differential geometry overview 1:37
 differential operators 4:319
 diffusion operators 3:6
 Dirac operator 2:74–86, 2:81
 dissipative operators 2:266–267
 evolution operators 2:336, 2:707
 b -Fourier integral operators 2:708
 Fredholm operators 1:92, 4:318
 and functional integration 2:440
 geometric operators 4:232
 Hecke operators 1:217
 hyponormal operators 4:641
 linear operators 4:633–645
 Møller operators 4:252
 and Newton problem 2:222
 normal operators 4:640
 b -pseudodifferential operators 2:701–712
 recursion operators 3:87–93, 3:91, 4:371–376
 scalar constraint operator 4:166
 unbounded operators 4:639
 and wavelets 5:435
 and Yang–Baxter equations 5:471
- Operator-valued functions, in axiomatic QFT 1:235
- Optical caustics
 codimension 3:626
 corank 3:626
 equivalence 3:625
 genericity 3:625
 and global asymptotics 3:625
 global organization 3:626
 in higher dimension spaces 3:627
 index 3:626
 as Lagrangian singularities 3:623
 and Maslov’s index 3:625
 modeling 3:622
 multiplicity 3:626
 and nonoptical caustics 3:627
 and optical catastrophes 3:623
 overview 3:620–627
 physical phenomenon 3:621
 as ray envelopes 3:622
 as short wave asymptotic 3:624
 stability 3:625
 stable, classification 3:626
 symmetrical caustics 3:627
 as wave front singularity locus 3:623
- Optical communications, nonlinear Schrödinger equations in 3:556
- Optics
 classical optics 2:88–89
 geometrical optics 1:635, 5:406
 nonlinear optics 3:552
 and nonlinear Schrödinger equations 3:552, 3:555
 nonoptical caustics 3:627
- Optimal cloning, quantum states
 asymmetric cloning 3:631
 covariant cloning maps 3:629
 distance measure 3:629
 finitely many states 3:631
 Gaussian states 3:631
 mathematical description 3:628
 overview 3:628–632
 phase-covariant pure-state cloning 3:630
 preferred state set 3:628
 property overview 3:629
 purification problems 3:632
 related problems 3:631
 solutions 3:629
 state estimation 3:630
 universal mixed-state cloning 3:631
 universal not 3:632
 universal pure-state cloning example 3:630

Optimal control, ordinary differential equations 1:638

Optimal transportation
Kantorovich problem 3:634
Monge's problem 3:634
overview 3:632–636
Wasserstein distance 3:636

Optimal transportation maps
with convex potentials 3:632
isoperimetric inequality proof 3:633
overview 3:632

Orbifolds
definition 1:382
fractional D3 branes 2:464

Orbits
basis sets 1:483
symbolic coding 2:725
and symplectic reduction 5:194

Orbit spaces, moduli spaces as 3:454

Order, in QFT 4:221

Order parameter, liquid crystals 3:322

Ordinary differential equations,
controllability 1:638

Ordinary random walks, into random
walks in random environments 4:370

Ordinary special functions
Bessel function 3:638
classical orthogonal polynomials 3:641,
3:643
confluent hypergeometric function 3:640
gamma and beta function 3:637
Gauss hypergeometric function 3:640
Hermite polynomials 3:642
hypergeometric series 3:639
Jacobi polynomials 3:643
Laguerre polynomials 3:642
modified Bessel function 3:639
overview 3:637–645

Orientability, vector bundles 1:492

Orientifolds, type IIB 5:107

Origin of universe, and mathematical
cosmology 1:657

Orthogonality relations
and Askey–Wilson polynomials 4:110
big q -Jacobi polynomial 4:111
Clifford algebras 1:520
on compact group 1:582
continuous q -ultraspherical
polynomials 4:111
 q -Hahn polynomials 4:112
Macdonald polynomials 4:113

Orthogonal space, Clifford algebras 1:524

Orthogonal wavelets, and multiresolution
analysis 5:429

Orthonormal bases, and quantum chaotic
system eigenfunctions 4:194

Oscillations, and multiscale
approaches 3:471–472

Oscillators, weakly coupled 5:448–453

OSFT *see* Open bosonic string field theory

P

Painlevé equations
Bäcklund and Miura
transformations 4:3
classical special solutions 4:3
Hamiltonians and Tau-functions 4:4
hierarchies 4:5
linear problems 4:2
overview 4:1–5
solution structure 4:2

Pairing, and fractional quantum Hall
effect 2:410

Pairs attractor–repeller, in general
dynamical systems 2:496

Pancharatnam connection, overview 2:529

Parabolic equations, for fluid
mechanics 2:367

Parabolic implosion, and holomorphic
dynamics 2:658

Parabolic methods, for constraint
equations 2:179

Parallel transport, and A–B
geometry 1:196

Parameterized post-Newtonian framework
bounds 2:485
and general relativity solar-system
tests 2:484

Parametric approximation, and multiscale
approaches 3:471

Parametric representation, in perturbation
theory 4:31

Parisi representation, and spin glass free
energy 4:662

Parseval formula, relative Fourier–Mukai
transforms 2:382

Partial decoupling, in effective field
theory 2:139–140

Partial differential equations
control 1:640
controllability via Coron return 1:641
integrable PDEs 1:346–351
and many-body problems 3:110
overview 4:6–7
quasiperiodic solutions 5:28
and scale spaces 3:6
Schrödinger equation control 1:641
separability in general setting 4:527
stochastic PDEs 5:69
vibrating string control 1:640

Partial hyperbolicity, and splitting 2:727

Particle collisions, and QFT analytic
program 4:469

Particle counters, in scattering
theory 4:460

Particle-like solutions, spherically
symmetric Einstein–Dirac–Yang–Mills
equations 2:71

Particle models,
supersymmetric 5:140–145

Particle physics
electroweak interactions 5:34
QFT applications 4:214
quantum chromodynamics 5:33
standard models overview 5:32–38

Particle systems
algebraic QFT 1:202
and asymptotic causality 4:479
and Fermionic systems 2:300
forward two-particle scattering
amplitude 2:88–89
 n -indistinguishable particles 2:444
infinite-particle systems 3:540
interacting particle systems 3:123–129
interacting stochastic particle
systems 3:130–135
limitations 4:205
massless particles 4:462
and Minkowski spacetime 1:102
multiparticle propagators 3:259
 n -particle quantum scattering
3:585–592
single-particle diffusion 3:256
spinning-particle mechanics 5:148
strongly damped Brownian
particles 5:87
three-particle quantum scattering
systems 3:588
two-particle boundstates 4:485
two-particle Lagrangian
dispersion 3:256
wave–particle dualism 2:713

Particle theories
one-particle Dirac theory 2:80
one-particle irreducible 4:31
Wigner's concept of particle 4:462

Partition functions
and abelian BF theories 1:261
in boundary conformal field
theory 1:336
BTZ black hole 4:501
and Chern–Simons theory 3:270
in combinatorics 1:554, 2:62
determinantal measures on 2:50
overview 4:347
zeroes, and Pirogov–Sinai theory 4:64

Parton distributions, and perturbative
quantum chromodynamics 4:150

Passively advected scalar, effective
diffusivity 3:481

Path integrals
Bismut–Chern character 4:10
covariant path integral 4:650
fermionic Brownian motion 4:11
Feynman path integrals 2:307–313, 4:9,
4:516
Gaussian path integrals 1:260
infinite-dimensional volume element 4:9
overview 4:8–12
and quantum dynamics 2:435

- Pathwise uniqueness, stochastic differential equations 5:68
- Pattern detection, wavelets applications 5:423
- Pauli representations, Clifford algebras 1:525
- Pauli–Villars regularization, and perturbation theory 4:33
- PCs *see* Periodic photonic crystals
- PCT symmetry, in axiomatic QFT 1:238
- P₁CT symmetry, in axiomatic QFT 1:238
- PDEs *see* Partial differential equations
- Peakons
- CH equation solution 4:16
 - generalization for Green’s functions 4:16
 - integrable dynamics of CH 4:15
 - and kinetic-energy norm 4:17
 - as mechanical systems 4:15
 - oblique wave front collisions 4:19
 - overview 4:12–20
 - reduction as momentum map 4:19
 - shallow-water background 4:13
 - singular solutions 4:14
- Peeling property
- and conformal infinity 1:224
 - Minkowski space stability 5:18
- Peierls condition, and Pirogov–Sinai theory 4:61
- Penrose inequality
- applications 2:519
 - in general relativity 2:513
 - overview 2:514
- Penrose transforms
- linear, zero-rest-mass free fields 5:313
 - and twistor theory 5:304
- Percolation theory
- contact processes 4:27
 - in cubic lattice subcritical regime 4:22
 - in cubic lattice supercritical regime 4:22
 - first-passage percolation 4:26
 - gradient percolation 4:26
 - invasion percolation 4:27
 - and lattices 4:25
 - near critical point 4:23
 - nonamenable graphs 4:26
 - overview 4:21–28
 - in two dimensions 4:24
- Perelomov construction, for coherent states 1:542–543
- Perfect fluids
- equations 2:196, 3:160
 - propagation speed and wave fronts 2:197
- Perfect-lens theorem
- generalized theorem 3:489
 - and geometries 3:490
 - and subwavelength imaging 3:488
- Perfect matchings, in combinatorics 1:568
- Perforated domains, and homogenization problems 2:452
- Perihelion advance, in general relativity solar system tests 2:484
- Periodic geodesics, and 2-step groups 4:98
- Periodic orbits
- bifurcations 4:592
 - bifurcation theory overview 1:285–290
 - in billiards 1:298
 - central manifolds 1:468
 - C¹-generic dynamical systems 2:499
 - and Hamiltonian system nonintegrability 2:626
 - linear systems 1:286
 - Malkin’s theorem 1:287
 - quasilinear systems 1:287
 - and semi-classical trace formula 4:513
- Periodic photonic crystals, and negative refraction 3:491
- Periodic points
- and holomorphic dynamics 2:654
 - in hyperbolic dynamical systems 2:726
- Periodic potential, effective diffusion in 3:480
- Periodic systems, overview 4:308
- Periods, Riemann surfaces 4:421
- Permanent stability, and Hamiltonian systems 2:633
- Persistent current, Ginzburg–Landau equations 2:549
- Perturbation of domains, and stochastic Loewner evolutions 5:85
- Perturbation theory
- classical gravitational many-body problems 2:578
 - clusters 2:58
 - in constructive QFT 1:619
 - dimensional regularization 4:33
 - examples and applications 2:59
 - and fermion fields 4:33
 - Feynman rules 4:30
 - graphs and trees 2:56
 - historical excursus 2:55
 - identity, and Leray–Schauder theory 3:282
 - infrared divergence 4:33
 - irreducible Green’s function 4:31
 - lattice regularization 4:33
 - loop integral computation 4:37
 - momentum cutoff regularization 4:32
 - multiscale analysis 2:58
 - overview 2:54–60
 - paradigmatic example 2:55
 - parametric representation 4:31
 - Pauli–Villars regularization 4:33
 - and quantum chromodynamics 4:35
 - and quantum electrodynamics 4:35
 - random, flows 4:331
 - string theory representation 4:39
 - techniques overview 4:28–41
 - in thermal QFT 5:234
 - Wick’s theorem 4:30
- Perturbative Chern–Simons–Witten theory, for finite-type invariant fundamental theorem 2:344
- Perturbative expansion
- operator product expansion 3:618
 - and quantum invariants 2:354
- Perturbative flow, and exact renormalization group 2:278
- Perturbative quantization, gauge theories 4:41
- Perturbative quantum chromodynamics, asymptotic freedom 4:149
- Perturbative renormalization
- and BRST, overview 4:41–47
 - characteristics 4:44
 - fermions 4:411
- Perturbing functions, classical mechanics overview 1:16
- PEs *see* Primitive equations
- Peter–Weyl theorem, and compact groups 1:582
- Pfaffian point processes, characteristics 2:53
- p*-forms
- definition 1:382
 - Lie algebra 1:147
 - nonabelian BF theories 1:258
- Phase coexistence
- geometry 1:74
 - and phase separation 1:75
- Phase-covariant pure-state cloning, quantum states 3:630
- Phase diagrams
- Falicov–Kimball model 2:287
 - and Pirogov–Sinai theory 4:60
- Phase-field method, for multicomponent fluid flow 3:140
- Phase locking, in periodic orbit bifurcations 1:286
- Phase separation
- and phase coexistence 1:75
 - and surface tension 1:75
- Phase space
- and algebraic QFT 1:201
 - coordinates, canonical transformations 1:4
 - diffusion 1:24
 - and Tomita–Takesaki modular theory applications 5:255
- Phase synchronization, chaotic systems 5:222
- Phase transition dynamics
- boundary condition dependence 4:54
 - in classical systems, overview 4:377
 - coarse graining transformations 4:55
 - competing species 4:55
 - conserved systems, spinoidal decomposition 4:49

- continuous system
 - overview 4:53–59
- continuum model 4:57
- contours and indicators 4:58
- equilibrium statistical mechanics
 - overview 1:59
- and gamma-convergence 2:454
- and gravitational collapse 1:671–672
- Ising model 4:55
- Kac potentials 4:57
- in lattice systems, overview 4:376
- nonconserved system phase
 - ordering 4:47
- nucleation 4:50
- overview 4:47–52
- and quantum chromodynamics 4:146
- quantum group symmetries 4:300
- in quantum systems 4:383
- and reflection positivity,
 - overview 4:376–386
- and statistical mechanics 4:53
- and symmetry breaking 1:70, 5:199, 5:203
- van der Waals liquid–vapor
 - transition 4:56
- Phase transitions
 - quantum phase transitions 4:289–295
- ϕ_{r-1} series
 - q -Saalschütz sum 4:108
 - summation and transformation
 - formulas 4:108
- Physical measures, random dynamical
 - systems 4:333, 4:336
- Piecewise-linear structures, for
 - manifolds 2:386
- Pilot-wave theory, overview 4:266
- Pin groups, Clifford algebras 1:529
- Pirogov–Sinai theory
 - applications 4:63
 - finite-size effects 4:64
 - generalizations 4:64
 - ground-state phase diagram 4:60
 - metastable free energy construction 4:62
 - overview 4:60–65
 - and Peierls condition 4:61
 - setting 4:60
 - zeros of partition functions 4:64
- Pi-stability, and derived categories 2:45
- Pitchfork bifurcation, and singularity
 - theory 4:590
- Planar three-body problem, Arnol'd's
 - theorem 3:194
- Plancherel measure
 - correlation functions 4:349
 - diagram dimension 4:348
 - dimension functions 4:350
 - limit shape and edge scaling 4:348
 - permutation enumeration 4:349
 - Schur functions and Cauchy
 - identity 4:349
- Planck-scale Hopf algebra, and
 - noncommutative spacetime 1:270
- Plane partitions, in combinatorics 1:570
- Planetary $(1+n)$ -body problem,
 - overview 3:189
- Plane waves, and Einstein equation
 - solutions 2:170
- Plasma, and negative refraction 3:484
- Plateau problem, for minimal
 - submanifolds 3:427
- PMP *see* Pontryagin maximum principle
- Poincaré algebra, for integrable
 - systems 2:428
- Poincaré center-focus problem, and
 - singularity theory 4:592
- Poincaré conjecture, and four-manifold
 - invariants 2:388
- Poincaré Denjoy theorem, and circle
 - diffeomorphisms 2:670
- Poincaré inequalities, in space
 - BV 2:412–413
- Poincaré integral
 - in multidimensional case 4:538
 - and separatrix splitting 4:536
- Poincaré invariance, and QFT analytic
 - program 4:468
- Poincaré rotation number, and circle
 - diffeomorphisms 2:669
- Point functions
 - analytic N -point functions 4:478
 - N -point functions 1:236, 5:231
 - one-point functions 2:162
- Point symmetry, in quasiperiodic
 - systems 4:312
- Point-vortex dynamics
 - evolution equations 4:67
 - Kirchhoff–Routh–Lin theory 4:75
 - Lin's special Green's function 4:76
 - method of images 4:73
 - N point vortices 4:73
 - overview 4:66–79
 - in 2D 5:396
 - vortex ring stability 4:71
- Pointwise in time, and viscous
 - incompressible fluids 5:372
- Poiseuille flow, in fluid
 - dynamics 1:281
- Poisson algebra, operads in 3:612–613
- Poisson brackets
 - from fluid Hamiltonian 2:598
 - and ideal Hamiltonian fluid 2:599
 - on symplectic manifold, deformations
 - Berezin-type star products 2:31
 - definitions 2:25
 - groups actions on star products 2:29
 - overview 2:24–34
 - star product classification 2:27
 - star product construction 2:26
 - star products on cotangent
 - bundles 2:32
- star products and function
 - subclasses 2:30
- Poisson equation
 - Dirichlet problem 2:221
 - Neumann problem 2:221
- Poisson groupoids
 - examples 3:316
 - geometry 3:314
 - Lie algebroids 3:319
 - overview 3:312–320
 - Poisson Lie groups as 3:317
- Poisson groups, and Lie bialgebras 1:512
- Poisson Lie groups
 - characteristics 1:515
 - definitions 3:316
 - integrable system applications 1:517
 - as Poisson groupoids 3:317
- Poisson–Lie T -duality, and
 - noncommutative spacetime 1:273
- Poisson manifolds
 - characteristics 4:79
 - cotangent bundles 3:317
 - deformation quantization 2:23
 - and Fedosov quantization 2:299
 - formal deformation quantization 2:19
 - and groupoids 3:315
 - Hamilton's equations on
 - cotangent bundles 3:38
 - finite-dimensional classical
 - mechanics 3:37
 - infinite-dimensional classical field
 - theory 3:37
 - overview 3:37
 - symmetry reduction 3:39
 - symplectic manifolds 3:38
- overview 1:511–518
- and quantum mechanics 2:2
- Poisson maps, and groupoids 3:315
- Poisson–Nijenhuis manifolds, integrable
 - systems on 3:88
- Poisson problem, in fluid mechanics 2:371
- Poisson quotients, and groupoids 3:315
- Poisson reduction
 - canonical Lie group actions 4:80
 - coisotropic submanifolds 4:81
 - cosymplectic submanifolds 4:82
 - Dirac's constraints formula 4:82
 - overview 4:79–84
 - by regular canonical distributions 4:81
 - and smooth functions 4:81
- Poisson relation, b -pseudodifferential
 - operator application 2:710
- Poisson–sigma model, and
 - Batalin–Vilkovisky formalism 1:252
- Poisson structures
 - and cotangent bundle reduction 1:665
 - and groupoids 3:315
 - and symplectic manifold 3:315
- Poisson tensor, and symplectic
 - leaves 4:80

- Poisson volume elements, and functional integration 2:438
- Polyakov measure, in Arakelov geometry 3:605
- Polygonal billiards
overview 4:84–87
periodic trajectories 4:85
rational polygons and flat surfaces 4:86
trajectory complexity 4:85
unfolding billiard trajectories 4:84
- Polylogarithms, in number theory 3:601
- Polymer activities, rigorous renormalization group analysis 2:279
- Polymeric liquid crystals, and constitutive equations 3:565
- Polymer models, abstract, and cluster expansion 1:532
- Polymer representations, rigorous renormalization group analysis 2:279
- Polynomiality, and finite-type invariants 2:346
- Polynomial-like maps, and complex renormalization 5:346
- Polynomial types
Askey–Wilson polynomials 4:110
big q -Jacobi polynomials 4:111
bracket polynomial 3:221
Charlier polynomials 3:643
classical knot polynomials 2:341
classical orthogonal polynomials 3:641
continuous Hahn polynomials 3:644
continuous q -ultraspherical polynomials 4:110
dual Hahn polynomials 3:644
four-manifold systems 2:389
G-invariant polynomials 2:324
 q -Hahn polynomials 4:112
Hall–Littlewood polynomials 4:114
Hermite polynomials 3:642
 q -hypergeometric orthogonal polynomials 4:109
Jack polynomials 4:114
Jacobi polynomials 3:643
Jones polynomial 3:179–187
knots and links 3:403
Krawtchouk polynomials 3:644
Laguerre polynomials 3:642
Landau polynomial 2:325
little q -Jacobi polynomials 4:111
Macdonald–Koornwinder polynomials 4:114
Macdonald polynomials 4:113, 4:114
Meixner–Pollaczek polynomials 3:644
Meixner polynomials 3:644
nondeterministic polynomial time 5:50
 q -Racah polynomials 4:111
Racah polynomials 3:645
Stieltjes–Wigert polynomials 4:112
and wavelets 5:433
Wilson polynomials 3:644–645
- Pontrjagin classes
index theorems 3:27
vector bundles 1:492
- Pontryagin maximum principle, and ordinary differential equation control 1:639
- Porous medium, effective diffusion in 3:480
- Position invariance, and Einstein equivalence principle 2:483
- Positive cones, in C^* -algebras 4:88
- Positive energy, and Tomita–Takesaki modular theory applications 5:255
- Positive integers, partitions, and q -special functions 4:116
- Positive maps, on C^* -algebras
in commutative algebras 4:89
completely positive maps 4:91
continuity and algebraic properties 4:89
morphisms 4:89
overview 4:88–94
positive-definite function on groups 4:89
Shur’s product of matrices 4:89
- Positive-mass theorem
and general relativity 2:512
and minimal surfaces 2:504, 3:429
in spinor calculus 4:672
- Positivity conditions
and $*$ -algebra construction 4:204
 N -point functions 1:237
and QFT analytic program 4:469, 4:473
- Positronium, quantum electrodynamics 4:173
- Postcritically finite maps, and Thurston’s theory 2:656
- Postselection, in classical weak measurement 4:277
- Potentials, electromagnetism overview 1:49
- Potthoff–Streit theorem, and noncommutative geometry 4:8
- Power counting, in constructive QFT 1:621
- PPN framework *see* Parameterized post-Newtonian framework
- PPTP *see* PPT-preserving operations
- PPT-preserving operations, and comparative entanglement 2:234
- Pressure
equilibrium statistical mechanics overview 1:52
and viscous incompressible fluids 5:377
- Primal–dual formulations, convex variational problems in duality 1:648
- Primary constraints, and Dirac algorithm 1:612
- Primitive equations
atmospheric 2:537
ocean 2:535
- Primordial background, and gravitational waves 2:584
- Principal bundles, characteristic classes 1:495
- Principal inertia moments, classical mechanics overview 1:11–12
- Principal symbol, on Lagrangian manifold 5:48
- Principle bundles, examples 1:488
- Probability, in dynamics 2:401
- Product Poisson structure, and groupoids 3:315
- Products matrices
 C^1 -generic diffeomorphisms 2:501
and random dynamical systems 4:332
- Projection methods
characteristic classes 1:493
and functional integration 2:437
Navier–Stokes condition 2:373
overview 3:473
- Projective modules, noncommutative tori 3:524
- Projective representations, definition 5:173–174
- Projective space
cohomology 1:551
tangent bundles 1:489
- Projective synchronization, chaotic systems 5:224
- Projective transformations, and Einstein–Cartan theory of gravitation 2:192
- Propagation, basic phenomena 5:402
- Propagation speed, perfect fluid 2:197
- Propagation of support, in symmetric-hyperbolic systems 5:162
- Propagators, classical mechanics overview 1:17
- PROPs, and operads 3:615
- Proton–proton scattering, quasi-dispersion relations 2:99
- Pseudogeometrical domain, atmospheric PEs 2:538
- Pseudo-Hermitian symmetric manifolds, characteristics 1:509
- Pseudohyperbolicity, and splitting 2:727
- Pseudorelativistic Schrödinger operators, properties 4:493
- Pseudo-Riemannian nilpotent Lie groups
Lorentzian groups 4:102
overview 4:94–103
2-step groups 4:96
- Pseudo-Riemann symmetric manifolds, overview 1:509
- Pullback
in K -theory 3:247
and twisted K -theory 3:252–253
- Pullback bundle, characteristics 1:490
- Pulsons
dispersionless b -equation solutions 4:16
overview 4:16

Pumped harmonic oscillator,
characteristics 4:161
Pure phases, statistical mechanics 5:56
Pure states, and comparative
entanglement 2:234
Purification, and quantum state
cloning 3:632
Purity, quantum dynamical
semigroups 4:163
Push-forward
in K -theory 3:247
in twisted K -theory 3:252–253

Q

Q -balls, characteristics 4:607
QCD *see* Quantum chromodynamics
QCI systems *see* Quantum completely
integrable systems
QC rigidity conjecture, and holomorphic
dynamics 2:657
QECCs *see* Quantum error-correcting
codes
QED *see* Quantum electrodynamics
QFT *see* Quantum field theory
QISM *see* Quantum inverse-scattering
method
QLs *see* Quadrilateral lattices
QM *see* Quantum mechanics
Quadratures, classical mechanics
overview 1:6
Quadrilateral lattices
Darboux-type transformations 3:82
discrete geometry 3:79
distinguished integrable reductions 3:83
Qualitative entanglement theory
pure states 2:230
setup 2:229
teleportation 2:231
witnesses and partial transpose 2:230
Quantitative transversality, and entropy
almost-transversality 2:240
characteristics 2:238
definitions 2:238
examples 2:240
as generic situation 2:239
measures and dimensions 2:239
overview 2:237–242
Quantization
abelian BF theories 1:260
Batalin–Vilkovisky
quantization 1:247–253
BRST quantization 1:386–392
canonical quantization 2:9, 2:541, 3:216
coherent states 1:544–545
constrained systems 1:615
deformation quantization 2:1–9, 2:9–16,
2:23, 2:24, 3:614
and Euclidean field theory 2:260
Fedosov quantization 2:291–300

formal deformation quantization 2:19
Kolmogorov–Arnol’d–Moser tori 3:584
open bosonic string field theory 5:98
perturbative quantization 4:41
quantum mechanics overview 1:116
and quantum N -body 4:305
and Schrödinger operators 4:489
Quantized universal enveloping algebras,
and deformation 2:19
Quantum affine algebras
definition 1:184
from quantum inverse-scattering
method 1:187
representations 1:185
 R -matrices 1:185
Quantum Calogero–Moser systems
excited states 4:126
factorized Hamiltonian 4:125
ground-state wave function and
energy 4:126
integrable deformations 4:128
Lax pair and conserved quantities 4:127
and Liouville integrability 4:123
overview 4:123–129
quantum *vs.* classical integrability 4:128
root systems 4:128
simplest cases 4:123
spectrum 4:126
spin models 4:128
superintegrability 4:127
triangularity 4:126
Quantum central-limit theorems
abnormal fluctuations 4:136
applications 4:138
CCR representations 4:132
lattice systems 4:131
linear response theory 4:139
normal fluctuations 4:133
overview 4:130–142
spontaneous symmetry breaking 4:140
Quantum channel capacity
and classical capacity 4:142–144
and coding theorems 1:428
elementary properties 1:426
entanglement-enhanced 1:418–423
and memory 1:430
overview 1:425
quantum channel-related 1:425
quantum information 1:424–431
related capacities 1:425
state-channel duality 1:427
variety 4:143
Quantum chaos
arithmetic quantum
chaos 1:212–221
and quantum ergodicity and
mixing 4:193
and random matrix theory 4:344
random waves and orthonormal
bases 4:194

Quantum chromodynamics
asymptotic freedom 4:147
background 4:145
chiral symmetry 4:147
confinement mechanisms 4:148
finite temperature 3:279
Green functions, phases, gauge
invariance 4:146
Lagrangian 4:145
low energies 2:144–145
numerical simulations 3:278
overview 3:275, 4:144–153
and perturbation theory 4:35
and random matrix theory 4:346
and renormalization group 4:406
scattering amplitudes 4:40–41
and Seiberg–Witten duality 4:509
at short and long distances 4:146
spontaneous breaking 4:147
in standard model 5:33
super-quantum chromodynamics 2:464
supersymmetric decomposition 4:40
and Yang–Mills theory 4:503
Quantum coding theorem, and information
systems 4:143
Quantum completely integrable systems,
eigenfunctions
Birkhoff normal forms 2:150
blow-up 2:149
lower bounds 2:151
modes and quasimodes 2:152
open problems 2:154
overview 2:148–155
QCI background 2:149
upper bounds 2:154
Quantum computing, and knot theory 3:227
Quantum conserved quantities,
Calogero–Moser systems 4:127
Quantum corrections
compactification 1:592
in gauge theory construction 1:365
Quantum cosmology
applications 4:155
general models 4:157
initial conditions 4:156
interpretations 4:157
minisuperspace approximation 4:154
overview 4:153–158
singularity applications 4:155
Quantum criticality, overview 4:292
Quantum cryptography, breaking 4:269
Quantum data compression
compression–decompression scheme and
fidelity 4:611
ergodic quantum sources 4:614
length 4:616
memoryless quantum sources 4:612
quantum sources and entropy 4:611
Quantum differential structures
classification 4:239

- Quantum differential structures
 - (*continued*)
 - covariant differential calculi 4:238
 - higher-order bicovariant calculi 4:238
 - Hopf algebra, preparations 4:238
 - Woronowicz theorems 4:238
 - Quantum dynamical semigroups
 - algebraic formalism 4:164
 - entropy and purity 4:163
 - ergodic properties 4:162
 - examples 4:160
 - and generators 4:159
 - Hamiltonian dynamics and semigroups 4:162
 - Itô–Schrödinger equations 4:163
 - nonlinear semigroups 4:164
 - overview 4:159–165
 - Quantum dynamics
 - in loop quantum gravity, overview 4:165–168
 - and path integrals 2:435
 - Quantum eigenstates, cat maps 1:219
 - Quantum electrodynamics
 - bound states 4:171
 - codifferential operator 1:142
 - current conversion 1:144
 - gauge group 1:148
 - gauge invariance 1:145
 - gauge potential 1:145
 - gauge potential and field strength 1:147
 - hydrogen-like atoms 4:171
 - and indefinite metrics 3:17
 - lepton anomalous magnetic moments 4:174
 - Lie-algebra-valued p -forms 1:147
 - Lorentz force equation 1:142
 - Maxwell equations 1:144
 - muonium 4:173
 - nonabelian gauge theories 1:146
 - origin 4:168
 - overview 1:141, 4:168–176
 - and perturbation theory 4:35
 - positronium 4:173
 - radiative correction structure 4:170
 - relativistic kinematics 1:141
 - renormalization problems 4:214
 - spatial dynamics 2:109
 - stochastic collapse dynamics 4:271
 - Yang–Mills equation 1:148
 - Yang–Mills equation and electromagnetism 1:149
 - Quantum energy inequalities, in axiomatic QFT 1:239
 - Quantum entropy
 - asymptotical properties 4:181
 - entropy history 4:177
 - overview 4:177–183
 - statistical thermodynamics 4:180
 - von Neumann entropy 4:178
 - Quantum equilibrium
 - and H -theorem 4:267
 - and pilot-wave theory 4:267
 - Quantum ergodicity
 - asymptotics eigenfunctions 4:186
 - and chaos conjectures 4:193
 - eigenfunction problems 4:191
 - and geodesic flow 4:189
 - and mixing 4:185
 - and operator time and space averages 4:190
 - overview 4:183–196
 - rate 4:193
 - wave group and geodesic flow 4:184
 - Weyl law and local Weyl law 4:186
 - Quantum error-correcting codes
 - definition 4:196
 - and fault tolerance 4:199
 - Quantum error correction
 - CSS codes 4:198
 - and fault-tolerant measurement 4:200
 - overview 4:196–201
 - stabilizer codes 4:197
 - Quantum evolution, operator approximation 2:708
 - Quantum field theory
 - algebraic QFT 5:388
 - Anderson–Higgs–Kibble mechanism 4:225
 - axiomatic QFT 1:235, 3:20
 - boson transformation 4:224
 - completeness and bound states 4:475
 - conformal QFT 5:334
 - constructive QFT 1:617–631
 - in cosmology 4:214
 - as covariant functors 1:200
 - in curved spacetime, overview 4:202–212
 - and deformation quantization 2:7
 - and dispersion relations 2:94
 - Euclidean QFT 1:617, 2:256–265
 - formulation 4:399
 - functional integrals 2:446
 - gauge symmetry 5:181
 - gerbes overview 2:539–546
 - history 4:212
 - and Ising models 5:326
 - and Jones polynomial 3:185
 - local implementation of symmetry 5:181
 - lower spacetime dimensions, symmetries, overview 5:172–179
 - noncommutative QFT 3:517
 - off-shell dispersion relations 2:97
 - operator product expansion
 - conformal field theory 3:619
 - free field theory 3:617
 - interacting theories 3:617
 - overview 3:616–620
 - perturbative example 3:618
 - two dimensions 3:620
 - Ward identities 3:619
 - overview 4:212–216
 - in particle physics 4:214
 - relativistic QFT 2:91
 - renormalizable, and Hopf algebra structure 2:678–686
 - renormalization problems 4:214
 - Schwarz-type topological QFT 4:494–503
 - S -matrix symmetries 5:182
 - soliton QFT 4:580
 - spacetime symmetries 5:180
 - spontaneously broken symmetry 5:182
 - structure functions, and dispersion relations 2:96
 - symmetry and order 4:221
 - symmetry overview 5:172, 5:179–183
 - symmetry rearrangement 4:223
 - thermal QFT 5:227–235
 - 3D topological QFT 4:119
 - Tomita–Takesaki modular theory
 - applications 5:255
 - with topological defects, overview 4:221–229
 - topological QFTs 1:232–234, 2:118–125, 2:390, 3:227, 5:278–290
 - 2D QFT 5:329
 - as two-level theory 4:222
 - types 5:180
- Quantum field theory with indefinite metric
 - Euclidean stochastic equations 4:217
 - Hilbert space structure condition 4:216
 - and nontrivial scattering 4:219
 - overview 4:216–221
- Quantum free gas, and classical physics 4:302
- Quantum geometry
 - basic framework 4:230
 - and Big Bang 4:235
 - black hole applications 4:233
 - geometric operators 4:232
 - overview 4:230–236
 - and superstring theories 5:139
- Quantum gravity
 - Hamiltonian constraint 3:218
 - and knot invariants, overview 3:215–220
 - loop quantum gravity 3:227, 3:339–343, 4:165–168
 - and random matrix theory 4:346
 - 3D 4:646
- Quantum group principal bundles
 - automorphisms 4:242
 - connections and connection forms 4:241
 - Dirac q -monopole 4:242
 - gauge transformations 4:242
 - generalizations 4:241
 - and q -instanton 4:243

- and matter fields 4:242
- quantum homogeneous bundles 4:240
- quantum principle bundles 4:239
- quantum trivial bundles 4:240
- universal calculus case 4:240
- Quantum groups
 - affine quantum groups 1:183–190
 - bicrossproduct Poincaré groups 1:272
 - braided groups 2:694
 - q -deformation coordinate algebras 2:692
 - q -deformation enveloping algebras 2:690
 - differential structures on algebra 4:237
 - Hopf algebras 2:687
 - noncommutative geometry 2:699
 - overview 2:687–700
 - quasi-Hopf algebras 2:697
 - self-dual 2:693
 - q -special function applications 4:115
 - symmetries 4:299
- Quantum Hall effect
 - conductivity 4:248
 - fractional quantum Hall effect 2:402–411, 4:250
 - integral effect 4:249
 - and noncommutative geometry 4:247
 - overview 4:244–251
 - research developments 4:250
- Quantum information theory
 - capacity for 1:424–431
 - channels
 - Clifford channels 2:330
 - and completely positive maps 1:473
 - instruments 1:476
 - Jamiolkowski isomorphism 1:474
 - observables 1:476
 - overview 1:472–477
 - parameter-dependent operations 1:476
 - preparations 1:476
 - Steinespring theorem 1:474
 - under symmetry 1:475
 - systems, states, algebras 1:472
 - and classical information theory 4:142
- Quantum invariants
 - as element of Habiro ring 2:355
 - overview 2:352
 - and perturbative expansion 2:354
- Quantum inverse-scattering method
 - as affine algebra origin 1:187
 - and Yang–Baxter equations 5:468
- Quantum Ising chain, characteristics 4:290
- Quantum jumps, overview 2:530
- Quantum kinetic theory, and Boltzmann equation 1:310
- Quantum lattice systems, and quantum central-limit theorems 4:131
- Quantum Lax pair, Calogero–Moser systems 4:127
- Quantum links, characteristics 2:353
- Quantum mechanical scattering theory
 - long-range interactions 4:258
 - overview 4:251–259
 - scattering matrix 4:257
 - smooth method 4:255
 - time-dependent, and Møller operators 4:252
 - time-independent theory 4:253
 - trace-class method 4:254
- Quantum mechanics
 - continuous spontaneous localization model 4:270
 - EPR–Bohm and Bell’s theorem 4:262
 - formal structure 1:113
 - foundational overview 4:260–265
 - framework 4:260
 - hidden variables 1:128, 4:269
 - hidden variables and nonequilibrium 4:266
 - H -theorem 4:267
 - interpretation problem
 - approaches 1:126
 - Malliavin calculus applications 3:387
 - noncommutative geometry 3:512
 - overview 1:109–131, 4:265–276
 - pilot-wave theory 4:266
 - and h -pseudodifferential operators 2:701
 - quantization and dynamics 1:116
 - quantum nonlocality 4:263
 - quantum reality 4:261
 - reality in spin-1/2 systems 4:262
 - Schrödinger’s cat 4:264
 - spontaneous localization model 4:274
 - and star products 2:2
 - and subquantum information and computation 4:269
 - and subquantum measurement 4:268
 - supersymmetric QM 5:145–151
 - symmetries 5:179
 - Wick quantization 1:123
- Quantum mixing
 - and chaos conjectures 4:193
 - overview 4:183–196
 - rate 4:193
- Quantum N -body problem
 - overview 4:283–289
 - reduction by rotations 4:285
 - reduction by translations 4:284
 - second quantization 4:305
- Quantum nonequilibrium
 - experimental signature 4:270
 - and hidden variables 4:266
 - and nonlocal signaling 4:268
 - and pilot-wave theory 4:267
 - and subquantum measurement 4:268
- Quantum nonlocality, and quantum mechanics foundation 4:263
- Quantum nonrelativistic
 - Calogero–Moser–Sutherland systems, overview 1:406
- Quantum particles, Fermionic systems 2:300
- Quantum phase transitions
 - coupled dimer antiferromagnet 4:291
 - Ising chain 4:290
 - Landau–Ginzburg–Wilson method 4:294
 - overview 4:289–295
 - and quantum criticality 4:292
- Quantum planes
 - characteristics 2:694
 - and deformations 2:19
- Quantum postselection, in weak measurement 4:280
- Quantum potential, equations 4:274
- Quantum reality
 - overview 4:261
 - in spin-1/2 systems 4:262
- Quantum relative entropy, properties 4:179
- Quantum relativistic
 - Calogero–Moser–Sutherland systems, overview 1:408
- Quantum reverse Shannon theorem, and entanglement-enhanced capacities 1:421
- Quantum Ruijsenaars–Schneider models, for integrable systems 2:429
- Quantum scattering
 - N -particle quantum scattering 3:585–592
 - relativistic, in linear response 2:94
- Quantum sources, and entropy, data compression 4:611
- Quantum spin chains, integrable, from affine quantum groups 1:189
- Quantum spin systems
 - associated mathematics 4:296
 - Hamiltonian dynamics 4:301
 - history 4:295
 - overview 4:295–301
 - phase transitions 4:300
 - source coding 4:615
 - symmetry and breaking 4:298
- Quantum states, optimal cloning
 - asymmetric cloning 3:631
 - covariant cloning maps 3:629
 - distance measure 3:629
 - finitely many states 3:631
 - Gaussian states 3:631
 - mathematical description 3:628
 - overview 3:628–632
 - phase-covariant pure-state cloning 3:630
 - preferred state set 3:628
 - property overview 3:629
 - purification problems 3:632

- Quantum states, optimal cloning
 - (*continued*)
 - related problems 3:631
 - solutions 3:629
 - state estimation 3:630
 - universal mixed-state cloning 3:631
 - universal not 3:632
 - universal pure-state cloning
 - example 3:630
 - Quantum statistics
 - Bose–Einstein condensation 1:82
 - and class field theory 3:603
 - Cooper pairs 1:83
 - Fermionic systems 2:300
 - Fermi surface 1:82
 - macroscopic system mathematical tools 4:306
 - ODLRO 1:83
 - overview 1:82, 4:302–308
 - problems and challenges 4:307
 - and quantum free gas 4:302
 - quantum N -body properties, second quantization 4:305
 - Quantum systems, reflection positivity and phase transitions
 - and Gaussian domination 4:384
 - nonabelian case 4:383
 - overview 4:383
 - Quantum three-manifolds
 - characteristics 2:353
 - closed invariants 4:119
 - modular category constructions 4:120
 - overview 4:117–122
 - ribbon and modular categories 4:117
 - 3D topological QFTs 4:119
 - TQFT integral structures 4:121
 - TQFT state-sum invariants 4:121
 - unitarity 4:120
 - Quantum time-continuous measurement, characteristics 4:281
 - Quantum unique ergodicity, and arithmetic quantum chaos 1:216
 - Quantum weak measurement
 - characteristics 4:278
 - classical measurement 4:277
 - and classical weak measurement 4:277
 - overview 4:276–283
 - postselection 4:280
 - related contexts 4:282
 - time-continuous
 - measurement 4:281
 - weak-value anomaly 4:280
 - Quantum weak mixing
 - overview 4:192
 - spectral measures and matrix elements 4:192
 - Quarks, lattice gauge theory 3:277
 - Quasars
 - microlensing characteristics 2:573
 - multiply-imaged 2:572
 - Quasiconvexity
 - and Hamiltonian system stability 2:633
 - and microstructures 5:368
 - Quasideterministic motion, and stochastic resonance 5:88
 - Quasi-dispersion relations, proton–proton scattering 2:99
 - Quasifree semigroups, characteristics 4:161
 - Quasi-Hopf algebras, and quantum groups 2:697
 - Quasi-invariance, heat measures on loop space 3:414
 - Quasilinear evolution equations, hyperbolic and parabolic 2:270
 - Quasilinear systems, periodic orbits 1:287
 - Quasiperiodic Fibonacci point set, properties 4:313
 - Quasiperiodic Fibonacci tiling, properties 4:313
 - Quasiperiodic functions, characteristics 4:310
 - Quasiperiodicity, classical mechanics
 - overview 1:7, 1:20, 1:26
 - Quasiperiodic point sets, characteristics 4:310
 - Quasiperiodic potentials, localization for and Cantor spectrum 3:338
 - dynamical localization 3:337
 - Lyapunov exponents 3:334
 - multidimensional case 3:336
 - by nonperturbative methods 3:336
 - overview 3:333–338
 - perturbative and nonperturbative approaches 3:334
 - without Lyapunov exponents 3:336
 - Quasiperiodic systems
 - overview 4:308–315
 - periodic system concepts 4:309
 - point sets and functions 4:310
 - point symmetry in 4:312
 - and scaling 4:311
 - Quasiperiodic tilings
 - fundamental domains 4:313
 - and windows 4:313
 - Quasispherical ansatz, for constraint equations 2:179
 - Quasistationary approximation, and multiscale approaches 3:472
 - Quasi-sure analysis, in Malliavin calculus 3:385
 - Quaternionic structures, Clifford algebras 1:519
 - Quaternionic Kähler manifolds, holonomy 2:186
 - Qubits, and entanglement 2:231
 - Queuing theory, and limit laws F_β 3:105
 - Quillen determinant
 - bundles 4:319
 - in finite dimensions 4:315
 - Fredholm determinant 4:316
 - Fredholm operators and determinant line bundles 4:318
 - overview 4:315–322
 - Quintic 3-fold
 - monodromy around LCLP 3:447
 - overview 3:446
 - Quivers
 - Lusztig’s variety 2:318
 - Nakajima’s quivers 2:320
 - overview 2:313–322
 - and path algebras 2:314
 - related topics 2:321
 - representations 2:315
 - QYBEs *see* Classical Yang–Baxter equations
- ## R
- Racah polynomials, characteristics 3:645
 - q -Racah polynomials, definition 4:111
 - Radiative corrections, structure 4:170
 - Radiofrequency pulses, and multidimensional nuclear magnetic resonance 3:597
 - Radon–Hurwitz numbers, Clifford algebras 1:527
 - Rado’s theorem, and minimal submanifolds 3:424
 - Rahman–Wilson biorthogonal rational functions, characteristics 4:112
 - Rains bound, properties 2:235–236
 - Ramond–Ramond states, definition 1:382
 - Randall–Sundrum brane worlds, characteristics 1:369
 - Random algebraic geometry
 - asymptotic counting formulas 4:327
 - attractor problem 4:325
 - elementary geometry 4:323
 - overview 4:323–329
 - Random bonds random walks in random environments, characteristics 4:365
 - Random dynamical systems
 - entropy formula measures 4:335
 - overview 4:330–338
 - perturbations of flows 4:331
 - physical measures as zero-noise limits 4:336
 - products of matrices 4:332
 - random maps 4:331
 - with random noise 4:330
 - stochastic stability of measures 4:333
 - Random-energy model, spin glasses 3:408–409
 - Random fields
 - determinantal random fields 2:47–54
 - Markov random fields 3:3
 - Random maps, dynamics 4:331

- Random matrix theory
 applications 4:346
 asymptotic freeness 2:421
 classical Gaussian ensembles 4:339
 correlation functions 4:340
 crossover transitions 4:343
 determinantal processes 2:586
 and determinantal random point fields 2:49
 in disordered systems 4:345
 ergodicity 4:341
 growth processes
 airy process 2:590
 basic model 2:587
 limit theorems 2:589
 overview 2:586–591
 Robinson–Schensted–Knuth correspondence 2:588
 variations 2:591
 integrable systems overview 3:102–106
 and limit laws F_β 3:103
 and Lyapunov exponents 3:351
 in many-body systems 4:344
 in mesoscopic systems 4:345
 overview 3:102, 4:338–347
 products 4:332
 in quantum chaos 4:344
 in quantum chromodynamics 4:346
 scattering systems 4:342
 statistical observables 4:340
 supersymmetry method
 approximations 5:158
 basic method 4:342
 bosonic variant 5:154
 determinants as Gaussian integrals 5:153
 fermionic variant 5:153
 Fyodorov’s method 5:154
 Gaussian ensembles 5:151
 Green’s functions 5:152
 nonlinear sigma model 5:159
 overview 5:151–160
 Riemannian symmetric superspace 5:158
 Schäfer–Wegner method 5:155
 supersymmetric variant 5:156
 zero-mode approximation 5:159
 symmetry classes
 Dirac fermions 5:211
 disordered superconductors 5:209
 Dyson’s threefold way 5:207
 framework 5:204
 overview 5:204–212
 symmetric spaces 5:205
 symmetry groups 5:204
 type I 5:206
 type II 5:205
 universality 3:103, 4:341
 wave functions 4:341
Random noise, in dynamics 4:330
Random partitions
 harmonic functions on Young graph 4:351
 overview 4:347–353
 Plancherel measure 4:348
Random perturbations
 flows 4:331
 metric entropy 4:335
Random point fields, determinantal 2:49
Random Schrödinger operators,
 properties 4:493
Random tilings, and limit laws F_β 3:104
Random walks
 balanced 4:369
 in continuous spontaneous localization model 4:273
 ordinary 4:354, 4:370
 and stochastic Loewner evolutions 5:84
 and weakly correlated diffusion 3:479
Random walks in random environments
 asymptotic velocity 4:356
 balanced random walks 4:369
 bounded jumps 4:363
 continuous-time 4:366
 critical exponent, excursion, traps 4:357
 definitions and equations 4:355
 diode model 4:362
 generalizations and variations 4:363
 in higher dimensions 4:367
 Kalikow’s condition 4:368
 limit distributions 4:359
 LLNs 4:368
 from ordinary random walks 4:370
 overview 4:353–371
 particle view 4:360
 random bonds 4:365
 Sinai’s localization 4:359
 Sznitman’s condition 4:368
 transience and recurrence 4:355
 zero-one laws 4:368
Random waves, and quantum chaotic
 system eigenfunctions 4:194
Rational conformal field theory, and
 braided monoidal categories 1:352
Rational function spaces, and holomorphic
 dynamics 2:656
Rational homology 3-spheres, and
 finite-type invariants 2:350
Rationality, Kontsevich integral
 construction 3:237
Rationalization, singularities 1:28
Rational polygons, and billiard
 dynamics 4:86
Ray envelopes, caustics as 3:622
Rayleigh–Bénard–Soret problem, and
 Newtonian fluids 3:499
Ray–Singer torsion, abelian BF
 theories 1:260
RBS problem *see* Rayleigh–Bénard–Soret
 problem
RCFTs *see* Rational conformal
 field theory
RCL *see* Reciprocal coupling limit
Real Clifford algebras,
 complexification 1:524
Real compact spin groups, Clifford
 algebras 1:530
Real structures, Clifford algebras 1:519
Reciprocal coupling limit, and EDYM
 equations 2:72
Reconstruction, and wavelet
 denoising 5:436
Rectifiability, and geometric measure
 theory 2:520
Rectifiable currents, and geometric
 measure theory 2:525
Rectifiable sets
 with finite measure 2:522
 and geometric measure theory 2:522
Recurrence
 in general dynamical systems 2:495
 C^1 -generic dynamical systems 2:500
Recursion operators
 and algebras of integrals of motion 3:91
 in classical mechanics,
 overview 4:371–376
 cotangent bundles 4:373
 cyclic systems in Levi–Civita 4:373
 on Jacobi manifolds 3:92
 Jacobi separable system 4:375
 on manifolds 3:87–93, 4:372
Redfield–Pólya theory, colored
 enumeration 1:558
Reduced Hamiltonian
 characteristics 2:614
 cosmological applications 2:622
Reductions
 asymptotic lattices 3:85
 Brown–Feynman reduction 4:38
 cotangent bundle reduction 1:658–667
 dimension reduction 2:453, 3:47
 distinguished integrable reductions 3:83
 Drinfeld–Sokolov reduction 1:295–296
 Hamiltonian reduction,
 overview 2:607–623
 integrable reductions 3:83
 Liapunov–Schmidt reduction 1:277
 Poisson reduction 4:79–84
 and QFT analytic program 4:469
 singular reduction 5:195
 symmetry reduction 3:39, 5:190–198
 symplectic reduction 3:73, 3:456, 4:514,
 5:190–198
Reductive Lie groups, overview 1:577
Reeb vector field, and contact
 manifolds 1:634
Reeh–Schlieder theorem, in axiomatic
 QFT 1:238
Reflection positivity
 in classical systems 4:377

- Reflection positivity (*continued*)
 and phase transitions,
 overview 4:376–386
 in quantum systems 4:383
- Reflection Yang–Baxter equations,
 characteristics 5:471
- Regularity
 minimal submanifolds 3:426
 second-order elliptic problems 2:224
 and semilinear wave equations 4:525
 and viscous incompressible fluids 5:373,
 5:377
 wavelet coefficients 5:434
- Regularization
 dynamical zeta functions,
 overview 4:386–391
 and minimal subtraction 4:402
 and perturbation theory 4:32
 singularities 1:28
- Regularized free field, overview 1:618
- Regular linear systems, on Riemann
 sphere 4:436
- Regular unimodal maps,
 characteristics 5:343–344
- Reidemeister moves, in knot theory 3:401
- Reissner–Nordström black hole
 definition 1:382
 and Einstein equation solutions 2:167
- Relations, topologies 1:138
- Relative entropy method, and interacting
 particle systems 3:128
- Relative equilibria
 and bifurcations 5:189
 and equivariant dynamics 5:185
- Relativistic bosons, field algebras and
 quasifree representations 1:322
- Relativistic Calogero–Moser systems, and
 sine-Gordon equation 4:582
- Relativistic fermions, field algebras and
 quasifree representations 1:322
- Relativistic kinematics, overview 1:141
- Relativistic local current algebra, for
 hadrons 1:674
- Relativistic quantum field theory
 and dispersion relations 2:91, 2:95
 scattering theory 2:94, 4:40, 4:251–259,
 4:342, 4:456, 4:465–475, 4:492
- Relativistic quantum physics, dispersion
 relation 2:88–89
- Relativistic wave equations
 for arbitrary spin 4:396
 in curved spacetimes 4:395
 with higher spin fields,
 overview 4:391–398
 for low spin in Minkowski
 spacetime 4:392
- Relativity
 Galilean relativity 1:44
 general *see* General relativity theory
 and Minkowski spacetime 1:101
 in Newtonian dynamics 1:41
 special relativity theory 3:503–509
- Relaxation, and microstructures 5:365
- REM *see* Random-energy model
- Renormalizable quantum field theory, and
 Hopf algebra structure
 graphs 2:680
 and Hochschild cohomology 2:684
 overview 2:678–686
 physical parameter
 diffeomorphisms 2:684
- Renormalization
 anomalies 4:405
 characteristics 4:404
 complete formulation 4:401
 complex, and universality 5:346
 and coordinate space 4:402
 and density of hyperbolicity 5:347
 and dimensional regularization and
 minimal subtraction 4:402
 and Feigenbaum Julia set
 geometry 5:347
 general counter-term
 formulation 4:403
 lattices 2:281
 and Mandelbrot set is locally connected
 conjecture 5:347
 Markovian diffusion 3:479
 and motivic Galois theory 3:606
 as multiscale approach 3:465–466
 one-loop calculations 4:399
 overview 4:399–407, 4:407–414
 in QFT 4:214, 4:399
 schemes 4:405
 self-avoiding walks 3:480
 and symmetries 4:404
 theorem overview 5:347
 in universality, overview 5:343–349
- Renormalization group
 applications and generalizations 4:406
 block spin RG 4:408
 bosons 4:414
 and condensed matter, overview 4:407
 as continuous semigroup 2:277
 as discrete semigroup 2:273
 equilibrium statistical mechanics
 overview 1:79
 fermions 4:411
 field theory and statistical
 mechanics 4:410
 mathematical structure 4:410
 and multiscale approach 3:476
 overview 2:272–281
 rigorous analysis overview 2:279
 scale decomposition RG 4:409
- Renormalization horseshoe
 and stochasticity 5:348
 and universality phenomenon 5:345
- Replica symmetry breaking, short-range
 spin glasses 4:572–573
- Representations
 Clifford algebras
 inductive construction 1:527
 overview 1:521
 types 1:525
 compact groups 1:581, 1:584
 Euler and Thom classes 3:391
 finite-dimensional, compact
 groups 1:584
 for finite Weyl system
 representation 2:332
 quantum affine algebras 1:185
 quivers 2:315
 in thermal QFT 5:230
- Representation theory
 basic Lie superalgebras 3:309
 and deformation quantization,
 overview 2:9–16
- Rescaled quantity, and multiscale
 approach 3:476
- Residue formula, and Hamiltonian group
 actions 2:605
- Resolvent equations, for three-particle
 quantum scattering systems 3:588
- Resolvent sets, and spectral theory
 4:635
- Resonance-free regions,
 characteristics 4:416
- Resonance geometry, and Hamiltonian
 system stability 2:632
- Resonances
 classical mechanics overview 1:21
 definition 4:416
 location 4:418
 overview 4:415–419
 and Painlevé equations 4:2
 resonance-free regions 4:416
 and time asymptotics 4:417
 and trace formulas 4:418
- Response kernels, scattering kernels
 as 2:94
- Restricted three-body problem
 classical gravitational many-body
 problems 2:576
 stability problems 5:20
- Restriction, stochastic Loewner
 evolutions 5:82
- Retarded kernels, and microcausality 2:95
- Reversed triangle inequality, in Minkowski
 spacetime 1:99
- Reversibility, and ergodic theory 2:254
- Reynolds equations, in turbulence
 theories 5:298
- Reynolds number
 in turbulence theories 5:297
 and viscous incompressible fluids 5:370
- RG *see* Renormalization group
- RH method *see* Riemann–Hilbert methods
- Rhombus tilings, in combinatorics 1:570
- RH problem, and soliton theory 3:95–96

- Ribbon categories, basic definitions 1:354
- Ricci curvature, differential geometry
overview 1:38
- Ricci flow
finite extinction for 3:431
singularity overview 4:584–588
- Riemann–Cartan geometry,
characteristics 2:192
- Riemann curvature, and geodesic
deviation 2:491
- Riemann–Hilbert methods
Bolibrukh’s method 4:439
and Deligne–Simpson problem 4:440
formulation and history 4:437
and holonomic quantum fields 2:663
integrable evolution equations 1:348
in integrable systems,
overview 4:429–435
Levelt’s result 4:439
linear systems on Riemann sphere 4:436
and noncommutative
spacetime 1:273–274
overview 4:436–441
setting 4:438
- Riemannian geometry, and contact
manifolds 1:634
- Riemannian holonomy groups
Berger’s classification 4:442
overview 4:441–446
- Riemannian manifolds
Hamilton–Jacobi equation on 4:527
Hodge decomposition 1:550
quantum completely integrable system
modes 2:152
semiclassical Schrödinger
operators 4:513
sharp Sobolev inequalities on 3:36
and Sobolev space inequalities 3:34
- Riemannian symmetric manifolds
Cartan domains 1:505
rank 1 1:505
- Riemannian symmetric spaces, and Lie
groups 3:301
- Riemannian symmetric superspace,
random matrix theory supersymmetry
method 5:158
- Riemann Penrose inequality
applications 2:519
in general relativity 2:513
overview 2:514
- Riemann–Roch theorems
and intersection theory 3:153
and Riemann surfaces 4:424
- Riemann sphere, linear systems on 4:436
- Riemann surfaces
basic definitions 4:420
Brill–Noether theory 4:426
compact Riemann surface moduli 4:427
compact Riemann surfaces 4:421
delta-bar problem on 2:40
divisors, linear systems, line
bundles 4:422
and Green’s conjecture 4:426
overview 4:419–429
periods and Jacobian 4:421
and Riemann–Roch theorem 4:424
uniformization 4:420
Verlinde formula and conformal
blocks 4:428
- Riesz potentials, Sobolev spaces 2:218
- Rigid body
classical mechanics overview 1:11
and Lie–Poisson bracket 3:39
and symmetry reduction 3:39
- Rigidity
map classes 3:436
and stationary black hole solutions 5:43
- Rigid monoidal categories, basic
definitions 1:352
- Rigid vertex graphs, and Vassiliev
invariants 3:223
- Ringed spaces, in deformation theory 2:17
- Ringel–Hall algebras, overview 2:317
- Rings, in topologies 1:139
- r -matrices
classical r -matrices 1:511–518
quantum affine algebras 1:185
and symplectic reduction 3:73
- RMT *see* Random matrix theory
- Robinson–Schensted–Knuth
correspondence, and random matrix
theory 2:588
- Robinson–Trautman solutions, Einstein
equations 2:171
- Rod climbing, as non-Newtonian
behavior 3:561
- Rogers–Ramanujan identities
overview 4:109
summation formula 4:109
- Ronkin function, and dimer problems 2:64
- Rooted trees, Hopf algebra 2:685
- Root lattices
canonical tilings 4:313
quasiperiodic point symmetry 4:312
- Root spaces, overview 3:293
- Root systems
constant term identities 4:112
 η -function identities 4:112
Macdonald–Koornwinder
polynomials 4:114
Macdonald polynomials 4:113
quantum Calogero–Moser systems 4:128
- Rotations, and quantum n -body
problem 4:285
- Rough paths, and stochastic differential
equations 5:70
- Rozansky’s Gaussian integral,
definition 2:354
- Rozansky–Witten theory, for finite-type
invariants 2:346
- RPC3BP, and Kolmogorov’s
theorem 3:191
- RSB *see* Replica symmetry breaking
- RS brane worlds *see* Randall–Sundrum
brane worlds
- RSK correspondence *see*
Robinson–Schensted–Knuth
correspondence
- R -spaces, noncompact Riemannian
symmetric manifolds 1:508
- RSW theory, and percolation in two
dimensions 4:24
- Ruijsenaars’ gamma function, elliptic
analog 4:115
- RWREs *see* Random walks in random
environments
- RYBEs *see* Reflection Yang–Baxter
equations
- ## S
- S'' , definition 1:382
- q -Saalschütz sum, for terminating balanced
 $3\phi_2$ 4:108
- Saddle-point manifold, random matrix
theory supersymmetry method 5:158
- Saddle-point problems
applications 4:450
details 4:449
examples 4:449
minimax 4:448
overview 4:447–456
superlinear problems 4:454
weak linking 4:454
- 3-Sasakian structures, description 2:186
- SAT *see* Satisfiability
- Satisfiability, and spin glass models 5:52
- Sato–Miwa–Jimbo theory, and holonomic
quantum fields 2:660
- Sato’s theory, and Toda lattices 5:242
- SC *see* Superconductivity
- Scalar constraint operator, in loop
quantum gravity 4:166
- Scale decomposition renormalization
group, setup 4:409
- Scale-echoing, and gravitational
collapse 1:668
- Scale spaces
basic principles 3:5
diffusion operators from 3:6
and partial differential equations 3:6
- Scaling limits
in dimer problems 2:65
and percolation in two dimensions 4:24
- Scaling theory
and perturbative quantum
chromodynamics 4:150
Plancherel measure 4:348
and quasiperiodicity 4:311
and 2D Ising model 5:324

- Scattering kernels
 - analytic structure nontriviality 4:470
 - and QFT analytic program 4:469
 - as response kernels 2:94
- Scattering matrix, and quantum
 - mechanical scattering theory 4:257
- Scattering theory
 - amplitudes in twistor space 4:40
 - analytic program 4:465–475
 - and asymptotic completeness 4:464
 - asymptotic particle counters 4:460
 - Haag–Ruelle theory 4:457
 - in linear response 2:94
 - LSZ formalism 4:459
 - and massless particles 4:462
 - mathematical setting and physical
 - motivation 4:456
 - physical motivation 4:456
 - quantum mechanical scattering
 - theory 4:251–259
 - in random matrix theory 4:342
 - in relativistic QFT, analytic program
 - overview 4:465–475
 - and Schrödinger operators 4:492
 - and Wigner’s concept of particle 4:462
- SCET *see* Soft-collinear effective theory
- SCFTs *see* Superconformal field theory
- Schäfer–Wegner method, random matrix
 - theory supersymmetry method 5:155
- Schlesinger’s equation
 - and isomonodromic
 - deformations 3:173
 - overview 2:20
- Schouten bracket, Lie algebroids 3:318
- Schouten–Nijenhuis bracket,
 - overview 2:562
- Schröder path, in combinatorics 1:555
- Schrödinger cocycles, and Lyapunov
 - exponents 3:352
- Schrödinger equations
 - and Aharonov–Bohm effect 1:193
 - classical mechanics overview 1:14
 - control 1:641
 - C-valued function 1:318–319
 - discrete 2D equation 3:86
 - and functional integration 2:440
 - inverse scattering for 3:159
 - Itô–Schrödinger equations 4:163
 - kinetic equation derivation 3:205
 - and linear evolution
 - equations 2:266–267
 - nonlinear equations 1:349, 3:552–560
- Schrödinger operators
 - eigenfunction properties 4:488
 - Lieb–Thirring inequalities 4:490
 - magnetic operators 4:490
 - N -body operators 4:491
 - and Novikov–Veselov hierarchy 2:37
 - overview 4:487–494
 - pseudorelativistic operators 4:493
 - quantization and semiclassical
 - limit 4:489
 - random operators 4:493
 - on Riemannian manifolds 4:513
 - and scattering theory 4:492
 - self-adjointness 4:487
 - and spectral analysis 4:488
 - and Sturm–Liouville theory 4:489
 - and theoretic physics 4:487
 - 2D, and Novikov–Veselov
 - hierarchy 2:37
- Schrödinger’s cat, and quantum mechanics
 - foundation 4:264
- Schrödinger systems, characteristics
 - 3:558
- Schumacher’s theorem, for memoryless
 - quantum sources 4:612
- Schur functions
 - Macdonald polynomials 4:114
 - and Plancherel measure 4:349
- Schwarzschild metric
 - and Einstein equation solutions 2:167
 - in general relativity 2:513
 - for stationary black holes 5:38–39
- Schwarzschild spacetime,
 - characteristics 4:210
- Schwarz-type topological quantum field
 - theory
 - Chern–Simons theory manifold
 - invariants 4:499
 - nonabelian Chern–Simons theory 4:496
 - overview 4:494–503
 - 3D Chern–Simons theory 4:495
 - 3D gravity and Chern–Simons
 - theory 4:501
- Sciama–Kibble field equations, and
 - Einstein–Cartan theory of
 - gravitation 2:193
- Scott’s conjecture, and \hbar -pseudodifferential
 - operators 2:703
- SCTF *see* Semiclassical trace formula
- Sears’ transformation, terminating
 - balanced ${}_4\phi_3$ 4:109
- Secondary constraints, and Dirac
 - algorithm 1:612
- Second-order elliptic problems, and linear
 - theory 2:223
- Sectors, algebraic QFT 1:202
- Secular divergences, as multiscale
 - approach 3:466
- Secure communication, chaos
 - synchronization applications 5:225
- Seiberg dualities, and gauge theory
 - classical theories 4:506
 - low-energy theories 4:504
 - Yang–Mills theory and QCD 4:503
- Seiberg–Witten duality, and gauge theory
 - low-energy solution 4:508
 - and QCD 4:509
 - Yang–Mills theory 4:507
- Seiberg–Witten equations, and four-
 - manifolds 2:391
- Seiberg–Witten Floer homology,
 - overview 2:363
- Seiberg–Witten invariants
 - in gauge theory 2:478
 - and monopoles 2:461
- Seiberg–Witten limit, and open-string
 - system 3:517
- Seiberg–Witten map, and noncommutative
 - gauge theory 3:520
- Seiberg–Witten theory
 - classical gauge theories 4:506
 - low-energy theories 4:504
 - overview 4:503–512
 - Yang–Mills theory and QCD 4:503
- Seifert manifolds, and Hamiltonian
 - reduction of Einstein’s
 - equations 2:609
- Selberg trace formula
 - and arithmetic quantum chaos 1:214
 - overview 4:513
- Self-adjoint generators, in linear evolution
 - equations 2:266–267
- Self-adjointness
 - and \hbar -pseudodifferential
 - calculus 2:705
 - and Schrödinger operators 4:487
- Self-avoiding walks, renormalization
 - analysis 3:480
- Self-critical phenomena,
 - characteristics 1:669
- Self-dual cones, and Tomita–Takesaki
 - modular theory 5:254
- Self-duality, Macdonald
 - polynomials 4:114
- Self-dual quantum groups,
 - characteristics 2:693
- Self-energy resummations, classical
 - mechanics overview 1:17
- Self-overlap, and short-range spin
 - glasses 4:573
- Self-similar fractals, and geometric measure
 - theory 2:521
- Self-similarity, and gravitational
 - collapse 1:670
- Semiboundedness, and \hbar -pseudodifferential
 - calculus 2:705
- Semicircle law, and free probability
 - theory 2:419
- Semiclassical approximation, and normal
 - forms 3:578–585
- Semiclassical expansions, and functional
 - integration 2:441
- Semiclassical mechanics, and \hbar -
 - pseudodifferential operators 2:702
- Semiclassical normal form, overview
 - 3:582
- Semiclassical Schrödinger operators, on
 - Riemannian manifolds 4:513

- Semiclassical trace formula, and closed orbits
- bifurcations 4:517
 - canonical measures and symplectic reduction 4:514
 - case with boundary 4:517
 - and eigenvalue statistics 4:517
 - Feynman integral derivation 4:516
 - integrable case 4:516
 - inverse spectra problem 4:517
 - maximally degenerated case 4:517
 - Newton flows 4:513
 - nondegeneracy 4:514
 - overview 4:512–518
 - periodic orbits 4:513
 - Selberg trace formula 4:513
 - semiclassical Schrödinger operators on Riemannian manifolds 4:513
 - smoothed density of states 4:515
 - spectral densities 4:515
 - Weyl expansion 4:516
- Semiconductor modeling, kinetic equations 3:205
- Semidirect sum Lie algebras, characteristics 3:290
- Semigroups
- continuous semigroups 2:277
 - discrete semigroup 2:273
 - functional analysis overview 1:93
 - heat semigroups 2:266–267
 - nonlinear dynamical semigroups 4:164
 - quantum dynamical semigroups 4:159–165
 - quasifree semigroups 4:161
- Semilinear wave equations
- classical results 4:518
 - global large waves 4:522
 - global small waves 4:523
 - low regularity 4:525
 - overview 4:518–526
 - Strichartz estimates 4:520
 - symmetries 4:520
- Semisimple Frobenius manifolds
- canonical coordinates 5:441
 - moduli 5:442
- Semisimple Lie algebras
- characteristics 3:291
 - and compact Lie groups 1:577
 - structure 3:293
- Semisimple Lie groups, and compact Lie groups 1:577
- Semisimple modular categories
- characteristics 1:356
 - conjugation properties 1:358
 - examples 1:359
 - Verlinde formula 1:357
- Sensitive dependence, in integrable dynamical systems 3:107
- Separable systems, Jacobi, in classical mechanics 4:375
- Separation of variables, for differential equations
- eigenvalue problem separation 4:533
 - Fourier method 4:526
 - Hamilton–Jacobi equation on Riemannian manifolds 4:527
 - Liouville integrable systems 4:534
 - overview 4:526–535
 - PDEs in general setting 4:527
- Separatrices, and Hamiltonian system nonintegrability 2:627
- Separatrix splitting
- and dynamics 4:539
 - exponentially small splitting 4:538
 - and geometric meaning of Melnikov integrals 4:536
 - and Melnikov integral 4:536
 - multidimensional case 4:537
 - overview 4:535–540
 - and Poincaré integral 4:536
 - and Poincaré integral in multidimensional case 4:538
 - standard example 4:537
- Serre spectral sequence, properties 4:630
- Set types
- Aubry sets 2:640
 - Fatou sets 2:654
 - Feigenbaum Julia sets 5:347
 - frequency set 2:709
 - Julia sets 2:400, 2:654
 - level-set methods 3:138
 - Mandelbrot set is locally connected conjecture 5:347
 - Mandelbrot sets 2:400, 2:657, 5:346
 - quasiperiodic Fibonacci point set 4:313
 - quasiperiodic point sets 4:310
 - rectifiable sets 2:522
 - resolvent sets 4:635
 - in topologies 1:138
- SFT *see* String field theory
- Shadowing, and hyperbolic dynamical systems 2:724
- Shale–Stinespring theorem, and fermions and bosons 1:325
- Shallow-water waves, for peakons 4:13
- Shannon coding theorem, and classical information theory 4:142
- Shannon’s noiseless channel coding theorem, and classical data compression 4:610
- Shape memory effect, and microstructures 5:363
- Shape-regular families, in fluid mechanics 2:369
- Shapiro time delay, in general relativity solar system tests 2:484
- Sharp Sobolev inequalities, on compact Riemannian manifolds 3:36
- Shear-dependent viscosity fluids, and constitutive equations 3:563
- Shear layers, in vortex flows 5:392
- Shear thickening, as non-Newtonian behavior 3:561
- Shear thinning, as non-Newtonian behavior 3:561
- Sheaves
- in open-string B model 2:41
 - and Stein manifolds 4:544
- Sherrington–Kirkpatrick model, mean-field spin glasses 3:407
- Shifted cotangent bundles, overview 2:564
- Shifting variables, random matrix theory supersymmetry method 5:157
- Shock matching, FRW–TOV, general theory 4:564
- Shock wave refinement
- FRW metric overview 4:559–570
 - FRW–TOV shock matching 4:565
- Shock waves, and symmetric–hyperbolic systems
- characteristics 5:163
 - Maxwell’s equations 5:164
 - overview 5:160–166
 - strong singularities 5:164
 - wave equation with variable coefficients 5:164
- Short-distance attraction, coalescence catastrophe from 1:84
- Short-range interactions, and spontaneous symmetry breaking 4:141
- Short-range spin glasses
- Edwards–Anderson Ising spin glass model 4:572
 - and mean-field models 4:665
 - open questions 4:574
 - overview 4:570–576
 - states and metastates 4:571
- Short wave asymptotic, caustics as 3:624
- SH systems *see* Symmetric–hyperbolic systems
- Shur’s product of matrices, and C^* -algebras 4:89
- Sigma models
- and Mathai–Quillen formalism 3:395
 - and solitons 4:604
- Signal compression, wavelets applications 5:424
- Signal-to-noise ratio, in magnetic resonance imaging 3:372
- Signal separation, wavelets applications 5:424
- Signature elliptic complex, characteristics 3:25
- Simons inequality, and minimal submanifolds 3:425
- Simple compact Lie groups, classification 1:578
- Simple covers, vector bundles 1:488
- Simple exclusion process, and thermodynamic fluctuations 3:365

- Simple harmonic oscillator, and deformation quantization 2:5
- Simple Lie algebras
characteristics 3:291
for Toda lattices 5:240
- Simple vectors, and geometric measure theory 2:523
- Simplicial homology, and chain complexes 1:546
- Simplicity
Clifford algebras 1:521, 1:525
graded algebras 1:523
- Sinai's localization, and random walks in random environments 4:359
- Sine-Gordon equation
classical solitons, antisolitons, breathers 4:578
integrable Hamiltonian system 4:577
and integrable QFT 3:55
overview 4:576–583
and relativistic Calogero–Moser systems 4:582
soliton QFT 4:580
- Single-particle diffusion, and Lagrangian dispersion 3:256
- Singular cohomology, and dual chain complexes 1:548
- Singular homology
basic properties 1:547
chain complexes 1:546
- Singular isothermal sphere, in gravitational lensing 2:570
- Singularity
bifurcation theory overview 4:588–594
classical gravitational many-body problems 2:578
deformation 2:18
fast–slow systems 4:593
functions 4:589
Hawking–Penrose singularity 3:347
homoclinic bifurcation 4:592
in intersection theory 3:154
inviscid compressible fluid flows 1:597
Lagrangian singularities 3:623
Landau singularity 4:479
normal forms 4:591
periodic orbit bifurcations 4:592
pitchfork bifurcation 4:590
Poincaré center-focus problem 4:592
quantum cosmology applications 4:155
rationalization and regularization 1:28
Ricci flow overview 4:584–588
in spacetime topology 4:617–623
and stability theories 4:592
and symmetric–hyperbolic systems 5:163
topological, and massless bosons 4:224
transcritical bifurcation 4:590
wave front singularity locus 3:623
wavelets applications 5:421
- Singular perturbations, as multiscale approaches 3:466, 3:470
- Singular point analysis, and Hamiltonian system integrability 2:629
- Singular reduction
cone spaces and local triviality 5:196
strata 5:195
stratified spaces 5:195
stratified spaces with smooth structure 5:196
symplectic stratification theorem 5:197
Whitney conditions 5:196
- Sinh–sine-Gordon theory, and classical integrable systems 3:51
- Site disordered models, mean-field spin glasses 3:408
- Six-vertex model, in combinatorics 1:572
- Skew products, and equivariant dynamics 5:185
- SK model *see* Sherrington–Kirkpatrick model
- Skyrmions, characteristics 4:604
- SLE *see* Stochastic Loewner evolution
- SL model *see* Spontaneous localization model
- Slow–fast systems, averaging in 1:230
- Slow–fast variables, and multiscale approaches
aggregation techniques and coarse-grainings 3:473
amplitude equations 3:471
averaging techniques 3:472
central manifold theorem 3:473
decomposition 3:471
numerical aspects 3:474
parametric approximation 3:471
projection techniques 3:473
quasistationary approximation 3:472
slow invariant manifolds 3:473
- Slow invariant manifolds, and multiscale approaches 3:473
- Slow motions, decoupling in averaging methods 1:228
- SM *see* Standard Model
- Smale horseshoe
and associated symbolic system 2:674
in dynamics 2:398
- Small coupling, even theory two-particle bound states 4:485
- Small-divisors problem, classical mechanics overview 1:21
- Small quantum fields, in constructive QFT 1:626
- S-matrix
and Bethe ansatz 5:466
exact, and affine quantum group symmetry 1:188
and integrable QFT 3:54
symmetries 5:182
- Smectic A phase, liquid crystals, elasticity 3:323
- Smectics, defects 3:327
- SMJ theory *see* Sato–Miwa–Jimbo theory
- Smooth dynamical systems,
division 2:721
- Smoothed density of states, and semi-classical spectra 4:515
- Smooth flows
incompressible perfect fluid 3:161
two-particle Lagrangian dispersion 3:256
- Smooth functions, and Poisson reduction 4:81
- Smoothing estimation, and viscous incompressible fluids 5:373
- Smooth manifolds
cohomology 1:549
de Rham cohomology 1:550
Grothendieck–Riemann–Roch theorem 3:247
- Smooth method, quantum mechanical scattering theory 4:255
- Smoothness, central manifolds 1:467
- Smooth solution, symmetric–hyperbolic systems 5:163
- Smooth symplectic manifolds, and Fedosov quantization 2:298
- SNR *see* Signal-to-noise ratio
- SNSs *see* Stationary nonequilibrium states
- Soap films, and minimal surfaces 3:430
- Sobolev embedding theorem, and inequalities 3:33
- Sobolev space
Bessel potentials 2:218
imbedding theorems 2:218
negative order 2:218
positive order 2:217
Riesz potentials 2:218
unbounded opens 2:219
weighted spaces 2:219
- Sobolev space inequalities
on compact Riemannian manifolds 3:36
and embedding theorem 3:33
isoperimetric and Euclidean type inequalities 3:35
and Kondrakov theorem 3:34
overview 3:32–36
and Riemannian manifolds 3:34
- Soft billiards, characteristics 2:720
- Soft-collinear effective theory,
overview 2:143
- Solar-system tests
Gravity Probe B 2:485
parameterized post-Newtonian framework 2:484
PPN parameter bounds 2:485
- Solenoids, as hyperbolic dynamical system example 2:725
- Solids, and capillary surfaces 1:436

- Solid–solid phase transformations, and microstructures 5:368
- Solid-state nuclear magnetic resonance, overview 3:599
- Solid states, valence bond 2:335
- Soliton ladder, and integrable PDEs 3:118
- Soliton quantum field theory, sine-Gordon equation 4:580
- Solitons
 - abelian Higgs vortices 4:605
 - analytical developments 3:95
 - bi-Hamiltonian methods
 - overview 1:290–296
 - characteristics 3:96
 - and delta-bar problem method
 - Davey–Stewartson-II and Ishimori-I equations 2:38
 - discrete systems 2:38
 - KP-II hierarchy 2:36
 - multidimensional problems 2:38
 - Novikov–Veselov hierarchy 2:37
 - and extended field configuration, overview 4:602–608
 - GMS solitons 3:521
 - Hamiltonian methods 1:291
 - history 4:595
 - and integrable nonlinear PDEs 3:115
 - and Kac–Moody Lie algebras
 - affine Lie algebras 4:596, 4:599
 - Bäcklund transformation 4:596
 - ball–box systems 4:601
 - Drinfeld–Sokolov method 4:599
 - equation hierarchy 4:597
 - geometric interpretation 4:596
 - Hamiltonian formalism 4:597
 - Hirota method 4:597
 - and infinitesimal transformation groups 4:598
 - integrable many-body problems 4:601
 - inverse-scattering method 4:595
 - Kadomtsev–Petviashvili hierarchy 4:598
 - Krichever’s work 4:600
 - Lax representation 4:595
 - overview 4:594–602
 - quasiperiodic solutions 4:600
 - Sato’s work 4:598
 - similarity solutions 4:601
 - special solutions 4:600
 - kinks and breathers 4:603
 - monopoles 4:606
 - Q-balls 4:607
 - sigma models 4:604
 - and sine-Gordon equation 4:578
 - Yang–Mills instantons 4:606
- Source coding
 - in classical data compression 4:609
 - quantum data compression, length 4:616
 - for quantum spin systems, overview 4:615
- Source-free equations, electromagnetism overview 1:48
- SoV *see* Separation of variables
- Space averages, and quantum ergodicity 4:190
- Space of rays, overview 2:528
- Spacetimes
 - anomalies in 1:210
 - anti-de Sitter–CFT
 - correspondence 1:174–183, 5:139
 - asymptotically flat 1:223
 - B-boundary 1:328
 - boundary from embeddings 1:326
 - boundary overview 1:326–333
 - causal boundary 1:329, 1:330, 1:332
 - curved spacetimes 2:69, 2:489, 4:202–212, 4:395
 - and Einstein equation solutions 2:166, 2:171
 - and interacting stochastic particle systems 3:133, 3:134
 - Kruskal–Szekeres spacetime 5:39
 - locality in BRST construction 1:390
 - lower dimensions QFT, symmetries 5:172–179
 - Minkowski spacetime 1:96–109
 - and Newton’s theory of gravitation 3:505
 - noncommutative spacetime 1:265–274, 2:141
 - symmetry characteristics 5:180
 - symmetry overview 5:166
 - 3 + 1 spacetime split 1:606
 - 2D spacetimes 3:348
- Spacetime topology
 - causality properties 4:621
 - null Raychaudhuri equation 4:620
 - overview 4:617–623
 - and singularity theorems 4:622
 - trapped surfaces 4:619
- Space types
 - arbitrary dimensions, integrable many-body problems 3:112
 - Banach space 3:331, 4:636
 - Calabi–Yau space 1:380–381
 - Cameron–Martin space 3:415–416
 - classical moduli space 1:258
 - classifying, cohomology 1:551
 - complex space deformation 2:18
 - cone spaces 5:196
 - coordinate space 4:402
 - energy–momentum space 2:95, 2:97, 4:468
 - Euclidean space 3:45
 - finite-element approximation space 2:368
 - Fock space 1:320–321, 1:539–540
 - functional spaces 1:135, 3:4
 - graded vector spaces 2:560
 - Hermitian symmetric spaces 2:31
 - higher dimension, caustics in 3:627
 - Hilbert space 4:216
 - homogeneous spaces 1:500–511
 - inner product spaces 1:519
 - instanton moduli space 3:48
 - kinematical Hilbert space 4:647
 - loop space 3:413–417
 - metric spaces 1:133
 - minisuperspace approximation 4:154
 - Minkowski space 1:326, 4:210, 5:14–19
 - moduli space 1:196, 2:470, 3:155, 3:449–458
 - Nambu space 5:209
 - nonspacelike cutlocus 3:344
 - normed space 3:282
 - operator spaces 1:135
 - orbit spaces 3:454
 - orthogonal space 1:524
 - phase space 1:4, 1:201, 5:255
 - rational function spaces 2:656
 - Riemannian symmetric spaces 3:301
 - Riemannian symmetric superspace 5:158
 - ringed spaces 2:17
 - root spaces 3:293
 - R-spaces 1:508
 - scale spaces 3:5
 - Sobolev space 2:217, 3:32–36
 - Sternberg spaces 1:666
 - stratified spaces 5:195
 - superspace 3:394, 5:124
 - symmetric spaces 5:205
 - topological space 1:89, 3:314
 - twistor space 4:40
 - 2D space 4:24
 - vector bundles 1:491
 - vector space 1:139, 1:523
 - wavelet space 5:410
 - weighted space 2:219
 - Weinstein space 1:666
 - whole space 2:225, 2:548
 - Wiener space 3:384
- Spatial chaos, dissipative dynamical systems 2:109
- Spatial discretization, and wavelet computation 5:417
- Spatial dynamics, dissipative dynamical systems 2:109
- Spatial three-body problem, and Arnol’d’s theorem 3:195
- Spatio-temporal chaos, dissipative dynamical systems 2:109
- Spatio-temporal synchronization, chaotic systems 5:224
- SPDEs *see* Stochastic partial differential equations
- Special functions, and algebraically completely integrable 3:76
- q*-Special functions
 - applications 4:115
 - conventions 4:105

- q -Special functions (*continued*)
 - elliptic hypergeometric series 4:115
 - q -hypergeometric orthogonal polynomials 4:109
 - overview 4:105–117
 - root system identities and functions 4:112
 - Special holonomy, and calibrated submanifolds 1:398
 - Special Lagrangian geometry
 - Calabi–Yau manifolds 1:399
 - compact SL m -fold deformations 1:400
 - compact SL m -fold obstructions 1:401
 - mirror symmetry 1:401
 - overview 1:399
 - submanifolds 1:400
 - Special Lagrangian manifolds
 - deformations 1:400
 - obstructions 1:401
 - Special Lagrangian submanifolds
 - characteristics 1:400
 - overview 1:398–402
 - Special relativity theory,
 - overview 3:503–509
 - Spectral action, noncommutative geometry 3:513
 - Spectral analysis
 - in axiomatic QFT 1:236
 - in fluid dynamics 1:282
 - and interacting stochastic particle systems 3:133
 - N -point functions 1:237
 - quantum Calogero–Moser systems 4:126
 - and Schrödinger operators 4:488
 - Spectral boundary conditions, manifolds 3:30
 - Spectral condition, and QFT analytic program 4:468
 - Spectral covers, and Fourier–Mukai transforms 2:383
 - Spectral decomposition, and hyperbolic dynamical systems 2:723
 - Spectral mapping theorem, overview 4:635
 - Spectral measures, and quantum weak mixing 4:192
 - Spectral problem, and wave propagation 1:342
 - Spectral resolution, and linear operators 4:642
 - Spectral sequences
 - derived functors 4:629
 - double complex 4:632
 - Eilenberg–Moore spectral sequence 4:632
 - exact couples 4:625
 - filtered chain complexes 4:625
 - filtered objects 4:624
 - graded, convergence 4:626
 - overview 4:623–633
 - Serre spectral sequence 4:630
 - standard examples 4:629
 - Spectral theory, for linear operators
 - adjoint operator 4:634
 - compact operators 4:638
 - complexification 4:637
 - essential spectrum 4:639
 - hyponormal operators 4:641
 - normal operators 4:640
 - operational calculus 4:636
 - overview 4:633–645
 - and resolvent sets 4:635
 - spectral resolution 4:642
 - unbounded operators 4:639
 - Spectral theory, functional analysis
 - overview 1:94
 - Spectral triples, noncommutative geometry 3:512
 - SP equations *see* Survey propagation equations
 - Spheres
 - finite-type invariants 2:347
 - hard spheres 1:308
 - isothermal sphere 2:580
 - nonsingular isothermal sphere 2:570
 - rational homology 3-spheres 2:350
 - Riemann sphere 4:436
 - singular isothermal sphere 2:570
 - topological sphere theorem 3:347
 - vector fields 1:528
 - Spherically symmetric
 - Einstein–Dirac–Yang–Mills equations
 - black-hole solution attempts 2:70
 - particle-like solutions 2:71
 - Spin
 - in axiomatic QFT 1:238
 - and Einstein–Cartan theory of gravitation 2:194
 - and fractional quantum Hall effect 2:410
 - Spin⁰ groups, Clifford algebras 1:529
 - Spin^c groups, Clifford algebras 1:529
 - Spin coefficients, overview 4:667–673, 4:670
 - Spin complex, index 3:28
 - Spin elliptic complex, characteristics 3:26
 - Spin ensemble, in NMR 3:593
 - Spin-exchange, and interacting stochastic particle systems 3:131
 - Spin-flip, and interacting stochastic particle systems 3:131
 - Spin foams
 - in 4D 4:651
 - overview 4:645–655
 - in 3D quantum gravity 4:646
 - Spin glasses
 - characteristics 2:553
 - derivation and interpolation formulas 4:659
 - diluted models 4:664
 - Edwards–Anderson Ising spin glass
 - model 4:572
 - future developments 4:666
 - Gaussian processes 3:408
 - Gibbs measures and multi-overlap distributions 3:409
 - mean-field model 4:657
 - model problem formulation 3:409
 - models and applications 3:409
 - overview 4:655–666
 - Parisi representation for free energy 4:662
 - and satisfiability 5:52
 - short-range and mean-field models 4:665
 - short-range spin glasses 4:570–576
 - thermodynamic limit and variational bounds 4:661
- Spin groups, Clifford algebras 1:529
- Spin Hamiltonian, in NMR 3:596
- Spin models
 - quantum Calogero–Moser systems 4:128
 - and Yang–Baxter equations 5:465, 5:470
- Spin network, and loop quantum gravity 3:340
- Spinning fluids, and Einstein–Cartan theory 2:193
- Spinning-particle mechanics,
 - overview 5:148
- Spinning top, and functional integration 2:445
- Spinoidal decomposition, in conserved systems 4:49
- Spinor algebra, in general relativity 4:668
- Spin-orbit problem
 - analytical results 5:24
 - numerical results 5:25
 - overview 5:24
- Spinor calculus, in general relativity 4:669
- Spinor groups, Clifford algebras 1:528
- Spinors
 - applications 4:673
 - overview 4:667–673
 - positive-mass theorem 4:672
 - spinor algebra 4:668
 - spinor calculus 4:669
- Spinor spaces, Clifford algebra
 - representations 1:526
- Spin relaxation, in NMR 3:595
- Spin structures, vector bundles 1:489, 1:492
- Spiral trajectories, in dynamics 2:400
- Splittings
 - dominated splittings 2:501, 2:726
 - and Hamiltonian system stability 2:635
 - hyperfine splitting 4:172
 - Lorentzian splitting theorem 3:346
 - separatrix splitting 4:535–540
 - 3 + 1 spacetime split 1:606

- Spontaneous breaking, quantum chromodynamics 4:147
- Spontaneous localization model, overview 4:274
- Spontaneously broken gauge theories, characteristics 5:201
- Spontaneously broken symmetry definition 5:173–174 and QFT 5:182
- Spontaneous symmetry breaking, characteristics 4:140, 5:168, 5:257
- Spurious pressure modes, in fluid mechanics 2:370
- Square lattices, scaling, and quasiperiodicity 4:311
- Squashed entanglement, properties 2:235–236
- Squeezing, branched manifolds 1:478
- SRT *see* Special relativity theory
- SSB *see* Spontaneous symmetry breaking
- Stability conditions asymptotic Lyapunov stability 5:4 asymptotic stability 5:188 bidirectional identical synchronization 5:219 C^* -algebras 3:250 in capillary surfaces 1:443 in celestial mechanics 5:20–25 and chaotic system synchronization 5:214 classical mechanics overview 1:20 equilibrium statistical mechanics overview 1:84 fluid flow 5:1–7 and hyperbolic dynamical systems 2:723 linear stability analysis 3:557 Lyapunov stability 3:350 nonlinear instability 5:4 in optical caustics 3:625 pi-stability 2:45 quasiperiodic motions 1:26 in random dynamical systems 4:333 resonances 1:21 restricted three-body problem 5:20 spin–orbit problem 5:24 ultraviolet stability 1:627
- Stability of matter first kind, proof 5:11 formulation 5:9 generalizations 5:10 overview 5:8–14 proof 5:12 and Thomas–Fermi theory 5:13
- Stability theory Arnol'd diffusion 4:592 and Hamiltonian systems averaging and exponential stability 2:631 overview 2:631–636, 4:592 permanent stability 2:633 and KAM Arnol'd diffusion and Nekhoroshev's theorem 5:29 lower-dimensional tori 5:28 overview 5:26–32 PDE quasiperiodic solutions 5:28 stability *vs.* chaos 5:30 Nekhoroshev theorem 4:592
- Stabilizer codes, and quantum error correction 4:197
- Stable ergodicity, and transitivity 2:727
- Stable manifolds, and hyperbolic dynamical systems 2:723
- Stable minimal surfaces, properties 3:426
- Stacks, and intersection theory 3:156
- Staggered approximations, in fluid mechanics 2:370
- Standard model chiral perturbation theory 2:145 as effective field theory 2:140 electroweak interactions 5:34 electroweak symmetry breaking 2:141 experimental fit 3:510 at low energies 2:144 and noncommutative spacetime 2:141 overview 3:509–515 particle physics overview 5:32–38 quantum chromodynamics 5:33 supersymmetric extensions 5:141 Yang–Mills–Higgs Ansatz 3:510
- Star products Berezin-type, and Hermitian symmetric spaces 2:31 on cotangent bundles 2:32 and deformation quantization 2:9 in Fedosov quantization 2:292 Morita equivalence 2:14 and quantum mechanics 2:2 on symplectic manifolds 2:26
- State-channel duality, and quantum channel capacity 1:427
- State summation, in knot theory 3:222
- Static bifurcation theory, overview 1:275
- Static electricity, electromagnetism overview 1:50
- Stationary black holes model-independent concepts 5:40 no-hairs theorems 5:43 overview 5:38–44
- Stationary nonequilibrium states, fluctuations 3:357–358
- Stationary phase approximation caustics 5:48 exact stationary phase 5:47 Lagrangian manifold principal symbol 5:48 method 5:47 overview 5:44–49 principle 5:46
- Stationary problems, and Boltzmann equation 1:310
- Statistical gravitational lensing, characteristics 2:575
- Statistical mechanics black holes, and branes, overview 1:373–382 and combinatorial problems, overview 5:50–55 constraint satisfaction problems 5:51 counting problem complexity 5:53 dynamical systems approach 3:540 equilibrium statistical mechanics 1:51–88 and gravitational collapse 1:671 interfaces overview 5:55–63 and Jones polynomial 3:184 message-passing algorithms 5:53 nonequilibrium statistical mechanics 3:530–540 numerical simulations 3:546 and phase transitions 4:53 quantum statistical mechanics 4:302–308 renormalization, overview 4:407–414 satisfiability and spin glass models 5:52 and Toeplitz determinants, overview 5:244–251
- Statistical observables, in random matrix theory 4:340
- Statistical systems, and integrable QFT 3:57
- Statistical thermodynamics, and quantum entropy 4:180
- Statistics algebraic QFT 1:202 bidimensional turbulence 5:77 in combinatorics 2:62 and limit laws F_β 3:104 quantum statistics 1:82, 2:300, 3:603, 4:302–308 and random matrix theory 4:346 and thermodynamic limit 2:65
- Steady parallel-flow model, Newtonian fluids 3:498
- Steenrod algebra, and chain complexes 1:548
- Steepness, and Hamiltonian systems 2:633
- Steinespring theorem completely positive maps 4:93 and quantum channels 1:474
- Stein manifolds classical types 1:502 complex analysis 4:543
- Sternberg spaces, symplectic leaves 1:666
- Stiefel–Whitney classes, vector bundles 1:491
- Stieltjes–Wigert polynomials, definition 4:112

- Stochastic collapse dynamics,
 requirements 4:271
- Stochastic differential equations
 distributional drift case 5:69
 existence in law 5:67
 Lipschitz case 5:65
 Malliavin calculus 5:70
 motivation and preliminaries 5:63
 non-semimartingales-driven SDEs 5:70
 numerical approximations 5:70
 overview 5:63–71
 pathwise uniqueness 5:68
 rough paths approach 5:70
 solutions overview 5:66
 and stochastic PDEs 5:69
 uniqueness in law 5:67
- Stochastic hydrodynamics
 Euler equations 5:73
 overview 5:71–79
 statistical solutions 5:71
 stochastic Navier–Stokes equations
 5:75
 vortices and bidimensional
 turbulence 5:77
- Stochasticity
 and renormalization
 horseshoe 5:348
 and topological dynamics 5:348
- Stochastic lattice gases, interacting
 systems 3:131–132
- Stochastic Loewner evolution
 in boundary conformal field
 theory 1:339
 and conformal field theory 5:83
 definition 5:80
 examples 5:84
 generalizations 5:85
 locality and restriction 5:82
 overview 5:80–86
- Stochastic maps, criteria 5:348
- Stochastic mechanics, overview 4:275
- Stochastic multiscale approaches,
 overview 3:471
- Stochastic Navier–Stokes equations, for
 hydrodynamics 5:75
- Stochastic partial differential equations,
 overview 5:69
- Stochastic resonance
 climate paradigm 5:86
 for continuously varying landscapes 5:91
 for landscapes 5:89
 overview 5:86–94
 quasideterministic motion transition
 criteria 5:88
 strongly damped Brownian particle 5:87
 and synchronization 5:93
- Stochastic stability, in random dynamical
 systems 4:333
- Stochastic water-wave models,
 overview 2:137
- Stokes equations
 basic problems 2:226
 in fluid mechanics 2:367, 2:372
 and geometric measure theory 2:524
- Stone’s theorem, and linear evolution
 equations 2:266–267
- Strain tensor, in fluid mechanics 2:365
- Strange attractors
 and chaos
 bounding tori 1:485
 bounding tori embeddings 1:487
 branched manifolds 1:478
 orbit basis sets 1:483
 overview 1:477–488
 and Lyapunov exponents 3:349–356,
 3:354
- Strata, singular reduction 5:195
- Stratified spaces
 singular reduction 5:195
 with smooth structure 5:196
- Stress tensor
 in conformal field theory 1:333
 in fluid mechanics 2:365
- Stress–tensor tensor, basic theory 4:208
- Stretching, branched manifolds 1:478
- Strichartz estimates, and semilinear wave
 equations 4:520
- String field theory
 closed bosonic SFT 5:100
 diagrams and minimal area metrics 5:99
 open bosonic SFT 5:95
 and operads 3:613
 overview 5:94–103
- String tension, lattice gauge theory 3:279
- String theory
 anomalies in 1:210
 black-hole solutions 1:374
 brane overview 1:375
 and BRST quantization 1:391
 and compactification 5:103
 and cosmology 5:110
 and D-branes 1:376, 3:522, 5:106
 and effective actions 5:109
 extended supersymmetric
 compactifications 1:590
 Fourier–Mukai transform
 overview 2:379–386
 gauge theories from,
 overview 2:463–468
 heterotic phenomenology 5:104
 higher-dimensional
 compactifications 1:590
 IIB, flux vacua 4:326
 instanton and monopoles 3:522
 intersecting brane worlds 5:107
 and matrix theory
 compactification 3:515
 moduli stabilization and supersymmetry
 breaking 5:109
 M-theory phenomenology 5:108
- noncommutative geometry
 overview 3:515–524
- NS branes 1:376
- open-string in B -field 3:515
- overview 5:103–110
- and random matrix theory 4:346
- representation 4:39
- and supergravity 5:127
- supersymmetric decomposition 4:40
- type II, duality, and twisted
 K -theory 3:253
- type IIB orientifolds 5:107
- String topology, homotopy and geometric
 perspectives 5:111–114
- Stringy corrections,
 compactification 1:592
- Strominger–Yau–Zaslow conjecture, and
 mirror symmetry 1:401, 3:443
- Strongly damped Brownian particles, and
 stochastic resonance 5:87
- Strong superadditivity, and entanglement
 measures 2:235
- Strong Szegő limit theorem, and Toeplitz
 determinants 5:245
- Structural glasses, characteristics 2:554
- Structure equations
 and asymptotic causality 4:486
 and irreducible kernels 4:484
- Structure functions
 and dispersion relations 2:96
 and turbulence 5:412
- Structure group, vector bundles
 lifting 1:489
 reduction 1:489
- Structure theorems, and gauge theoretic
 invariants of four-manifolds 2:459
- Sturm–Liouville theory, and Schrödinger
 operators 4:489
- Subadditivity method, equilibrium
 statistical mechanics overview 1:85
- Subdiffraction imaging,
 overview 3:483–491
- Subfactor theory, von Neumann algebras
 algebraic QFT 5:388
 analytic classification theory 5:387
 constructions and combinatorial
 classification 5:387
 overview 5:385–390
 topological invariants 5:388
- Subharmonic method, and Lyapunov
 exponents 3:353
- Submanifold types
 coisotropic submanifolds 3:315, 4:81
 Lagrangian submanifolds 3:315, 3:623
 minimal submanifolds 3:420–432
 special Lagrangian
 submanifolds 1:398–402
- Submaximal isotropy subgroups, and
 equivariant bifurcation theory 5:188
- Subquantum computation, overview 4:269

- Subquantum information, and quantum nonequilibrium 4:269
- Subquantum measurement, and quantum nonequilibrium 4:268
- Subwavelength imaging, and perfect-lens theorem 3:488
- Successive regimes, and multiscale approaches 3:469
- Sugawara algebra, overview 1:676
- Summation formulas
 - Jackson's ${}_8W_7$, elliptic analog 4:115
 - Rogers–Ramanujan identities 4:109
- Superadditivity, and entanglement measures 2:235
- Superalgebras, Lie type 3:305–311
- Supercentrality, graded algebras 1:523
- Superconductivity
 - bosons, and renormalization groups 4:414
 - and Ginzburg–Landau equations 2:550
- Superconductors
 - characteristics 2:306
 - disordered superconductors 5:209
 - and limit laws F_β 3:105
 - and topological defects 5:262
 - vortices 5:395
- Superconformal field theory
 - geometric setup 3:440
 - overview 3:439–448
- Superfluids
 - BCS–Leggett pair condensation 5:118
 - Bose–Einstein condensation 1:316, 5:117
 - boson and fermion degeneracy 5:116
 - condensates and excitation gas response 5:119
 - London quantum hydrodynamics 5:117
 - overview 5:115–121
 - vortices 5:395
- Supergravity
 - auxiliary fields and superspace 5:124
 - definition 1:382
 - extended 5:126
 - in higher dimensions 5:126
 - minimal action 5:123
 - minimal $D=4$ 5:122
 - $N_s=8$ 1:591
 - $N_s=16$ 1:590
 - $N_s=32$ 1:590
 - overview 5:122–128
 - strings and branes 5:127
- Superharmonic functions, in general relativity 2:512
- Superimposed regimes, and multiscale approaches 3:469
- Superintegrability, quantum
 - Calogero–Moser systems 4:127
- Super Lie algebras, characteristics 1:522
- Superlinear problems,
 - characteristics 4:454
- Supermanifolds
 - algebra basics 5:129
 - algebraic approach 5:131
 - calculus on 5:132
 - geometric approach 5:130
 - overview 5:128–133
- Supermassive stars, and gravitational waves 2:584–585
- Supernovas, and gravitational waves 2:584
- Super-quantum chromodynamics,
 - fractional D3 branes 2:464
- Super-renormalizability
 - characteristics 4:404
 - definition 4:482
- Superselection sectors, symmetry
 - connection 5:174
- Superselection theory, in 2D model 5:331
- Superspace
 - and Mathai–Quillen formalism 3:394
 - minisuperspace approximation 4:154
 - Riemannian symmetric superspace 5:158
 - and supergravity 5:124
- Superstring theory
 - Calabi–Yau compactification 1:588
 - compactification
 - general framework 1:586
 - $N=1$ supersymmetry in four dimensions 1:593
 - overview 1:586–595
 - stringy and quantum corrections 1:592
 - and D-branes 5:136
 - developments 5:139
 - dualities and M theory 5:137
 - five theories 5:133
 - interactions and effective theories 5:135
 - $N_s=8$ 1:591
 - $N_s=16$ supergravity 1:590
 - $N_s=32$ supergravity 1:590
 - overview 5:133–140
- Supersymmetric decomposition, and string theory 4:40
- Supersymmetric grand unified theories,
 - overview 5:143
- Supersymmetric oscillator
 - and classical supersymmetry 5:147
 - construction 5:148
 - overview 5:145
- Supersymmetric particle models,
 - overview 5:140–145
- Supersymmetric quantum mechanics
 - Mathai–Quillen formalism
 - applications 3:397
 - oscillator construction 5:148
 - overview 5:145–151
 - spinning-particle mechanics 5:148
 - Witten index 5:150
- Supersymmetry
 - anticommuting variables 5:146
 - breaking, and string theory 5:109
 - definition 1:382, 5:173–174
 - extended, twisting, and TQFT 5:281
 - extension 5:149
 - $N=1$, and compactification 1:593
 - properties 5:171
 - and random matrix theory
 - approximations 5:158
 - bosonic variant 5:154
 - determinants as Gaussian integrals 5:153
 - fermionic variant 5:153
 - Fyodorov's method 5:154
 - Gaussian ensembles 5:151
 - Green's functions 5:152
 - nonlinear sigma model 5:159
 - overview 4:342, 5:151–160
 - Riemannian symmetric superspace 5:158
 - Schäfer–Wegner method 5:155
 - supersymmetric variant 5:156
 - zero-mode approximation 5:159
 - in standard model extension 5:141
- Supertrace, in Mathai–Quillen construction 3:390
- Super Yang–Mills, and wrapped D5 branes 2:466
- Surface diffeomorphisms, wild behavior,
 - local C^2 -genericity 2:498
- Surfaces
 - area-minimizing surfaces 3:428
 - capillary surfaces 1:431–445
 - compact energy hypersurfaces 3:437
 - compact Riemann surfaces 4:421
 - crystalline surfaces 2:66
 - differential geometry overview 1:33
 - discrete integrable surfaces 3:84
 - Fermi surface 1:82, 4:411
 - flat surfaces 4:86
 - Friedman–Robertson–Walker surfaces 1:371
 - hyperbolic surfaces 1:213
 - Lorentz surfaces 3:348
 - marginally trapped surfaces 2:505
 - maximal surfaces 2:507
 - minimal surfaces 1:644, 2:503, 3:429
 - Riemann surfaces 4:419–429
 - stable minimal surfaces 3:426
 - trapped surfaces 4:619
- Surface tension
 - and dimer problems 2:64
 - and phase separation 1:75
 - statistical mechanics 5:56
- Survey propagation equations, and message-passing algorithms 5:53
- Symbolic coding, orbits and ergodic theory 2:725
- Symmetrical caustics, characteristics 3:627
- Symmetric attractors, and equivariant dynamics 5:186
- Symmetric classes
 - Dirac fermions 5:212

- Symmetric classes (*continued*)
 disordered superconductors 5:210
 Dyson's threefold way 5:207
- Symmetric cones, noncompact Riemannian
 symmetric manifolds 1:505
- Symmetric-hyperbolic systems
 Cauchy-Kowalewska systems 5:166
 compressible fluids 5:165
 conservation laws 5:161
 definitions 5:161
 Einstein's equations 5:165
 energy identity 5:162
 jump discontinuities 5:163
 modern context 5:160
 numerical scheme 5:163
 propagation of support 5:162
 and shock waves 5:160–166, 5:164
 smooth solution construction 5:163
 strictly hyperbolic equations 5:162
 Tricomi's equations 5:165
- Symmetric spaces, random matrix
 theory 5:205
- Symmetrization, in symmetric-hyperbolic
 systems 5:161
- Symmetrization maps,
 Le-Murakami-Ohtsuki
 invariant 2:351
- Symmetrized monomials, Macdonald
 polynomials 4:114
- Symmetry
 affine quantum group 1:188
 and $*$ -algebra construction 4:204
 algebraic QFT 1:202
 and Bäcklund transformations 1:245
 channels under 1:475
 and conservation laws
 baryon-antibaryon asymmetry 5:170
 baryon and lepton numbers 5:169
 discrete symmetries 5:169
 gauge theories 5:167
 general relativity 5:171
 overview 5:166–172
 spontaneous symmetry breaking 5:168
 supersymmetry 5:171
 topological changes 5:170
 dynamical rearrangement 4:223, 4:227
 in dynamical systems
 bifurcations 5:189
 equivariant bifurcation theory 5:187
 equivariant dynamics 5:185
 Hopf bifurcation and mode
 interactions 5:189
 isotropy subgroups and commuting
 linear maps 5:184
 overview 5:184–190
 in elementary quantum
 mechanics 5:179
 equilibrium statistical mechanics
 overview 1:65
 in Falicov-Kimball model 2:286
 in lower spacetime dimension QFT,
 overview 5:172–179
 in mathematical cosmological
 models 1:656
 in QFT 4:221, 5:172
 in QFT, algebraic aspects
 gauge symmetry 5:181
 local implementation 5:181
 overview 5:179–183
 S -matrix symmetries 5:182
 spacetime symmetries 5:180
 spontaneously broken symmetry 5:182
 quantum spin systems 4:298
 and renormalization 4:404
 and semilinear wave equations 4:520
 simple mechanical system
 reconstruction 1:663
 $SO(5)$, in antiferromagnetism–
 superconductivity unification 2:648
 spacetime symmetries 5:166
 superselection sector connection 5:174
 in topological QFT 1:232
- Symmetry breaking
 bifurcations 5:189
 in dynamical systems,
 overview 5:184–190
 effective field theory 2:139–140
 Elitzur's theorem 5:202
 equivariant bifurcation theory 5:187
 equivariant dynamics 5:185
 in field theory, overview 5:198–204
 finite groups 2:322–327
 global symmetries 5:198
 Goldstone bosons 5:200
 Hopf bifurcation and mode
 interactions 5:189
 isotropy subgroups and commuting
 linear maps 5:184
 phase transitions 1:70, 5:199
 phase transitions and crossovers 5:203
 spontaneously broken gauge
 theories 5:201
 and standard model 2:141
- Symmetry classes, in random matrix theory
 Dirac fermions 5:211
 disordered superconductors 5:209
 Dyson's threefold way 5:207
 framework 5:204
 overview 5:204–212
 symmetric spaces 5:205
 symmetry groups 5:204
 type I 5:206
 type II 5:205
- Symmetry groups, random matrix
 theory 5:204
- Symmetry reduction
 overview 5:190–198
 and Poisson manifolds 3:39
 system symmetries 5:191
 vector fields 5:191
- Symplectic diffeomorphisms, and
 Hamiltonian system stability 2:632
- Symplectic four-manifolds,
 characteristics 2:393
- Symplectic geometry, and contact
 manifolds 1:633
- Symplectic groupoids
 examples 3:316
 geometry 3:314
 Lie algebroids 3:319
 overview 3:312–320
- Symplectic leaves, Sternberg and Weinstein
 spaces 1:666
- Symplectic manifolds
 canonical Poisson structure 3:315
 formal deformation quantization 2:19
 graded 2:564
 and groupoids 3:314
 and Hamilton's equations 3:38
 integrable systems and recursion
 operators 3:87–93
- Poisson bracket deformations
 Berezin-type star products 2:31
 definitions 2:25
 groups actions on star products 2:29
 overview 2:24–34
 star product classification 2:27
 star product construction 2:26
 star products on cotangent
 bundles 2:32
 star products and function
 subclasses 2:30
- Symplectic orthogonality, and
 groupoids 3:315
- Symplectic quotients
 cohomology rings 2:605
 and Hamiltonian group actions 2:602
- Symplectic reduction
 basic theorem 5:193
 dynamics reconstruction 5:193
 momentum maps 5:192
 orbit formulation 5:194
 overview 5:190–198
 and r matrices 3:73
 and semiclassical trace formula 4:514
 vector bundles 3:456
- Symplectic stratification theorem, and
 singular reduction 5:197
- Symplectomorphism group, as infinite-
 dimensional Lie group example 3:42
- Synchronization
 and stochastic resonance 5:93
 and weakly coupled oscillators 5:451
- Synchronization, in chaotic systems
 anticipating synchronization 5:224
 applications 5:225
 bidirectional identical
 synchronization 5:218
 generalized synchronization 5:221
 identical synchronization 5:215

- lag synchronization 5:223
- overview 5:213–226
- phase synchronization 5:222
- projective synchronization 5:224
- spatio-temporal synchronization 5:224
- stability 5:214
- unidirectional identical
 - synchronization 5:217
- Synchronization manifold, and
 - bidirectional identical
 - synchronization 5:219
- SYZ conjecture *see*
 - Strominger–Yau–Zaslow conjecture
- Szegő limit theorem, and Toeplitz
 - determinants 5:245
- Sznitman’s condition, random walks in
 - random environments 4:368
- T**
- Tachyon condensation, and
 - D-branes 3:522
- Tangent bundles
 - Lie algebroids 3:317
 - Lie groupoids 3:314
 - projective space 1:489
- Tangent cone analysis, and minimal
 - submanifolds 3:427
- Tau-functions, and Painlevé equations 4:4
- Taylor–Couette problem, in fluid
 - dynamics 1:281
- Teleportation, and qualitative
 - entanglement theory 2:231
- Temperature
 - equilibrium statistical mechanics
 - overview 1:52, 1:67
 - and nonequilibrium statistical
 - mechanics 3:534
- Temperley–Lieb algebra, and Jones
 - polynomial 3:181
- Tensor products
 - construction, discrete wavelets 5:432
 - multiplicities, and compact groups 1:585
 - vector bundles 1:490
- Tensors, graded algebras 1:523
- Tetrad formalism, general relativity 1:415
- Theory of heaps, in combinatorics 1:562
- Thermal equilibrium, in thermal QFT
 - characterization 5:229
 - n -point functions 5:231
 - overview 5:228
 - particle aspects 5:232
 - state existence 5:230
 - thermal representations 5:230
- Thermal fields, asymptotic
 - dynamics 5:234
- Thermal gauge theory, from near-extremal
 - D3 branes 1:176
- Thermal model, in thermal QFT 5:233
- Thermal quantum field theory
 - asymptotic dynamics 5:234
 - free fields model 5:233
 - model 5:233
 - n -point functions 5:231
 - observables and states 5:228
 - overview 5:227–235
 - particle aspects 5:232
 - perturbation theory 5:234
 - representations 5:230
 - state characterization 5:229
 - state existence 5:230
 - thermal equilibrium overview 5:228
- Thermocapillary convection, and
 - Newtonian fluids 3:502
- Thermodynamic fluctuations
 - Boltzmann–Einstein formula 3:357
 - and dynamical Boltzmann–Einstein
 - formula 3:359
 - and Hamilton–Jacobi equation 3:360
 - and Kawasaki–Glauber dynamics
 - 3:365
 - models 3:364
 - nonequilibrium fluctuation dissipation
 - relation 3:361
 - and nonequilibrium statistical
 - mechanics 3:535
 - overview 3:357–366
 - simple exclusion process 3:365
 - and zero-range process 3:364
- Thermodynamic limits
 - adiabatic piston 1:165
 - and dimer problems 2:63
 - equilibrium statistical mechanics
 - overview 1:57, 1:69
 - multiscale approaches 3:470–471
 - spin glasses 4:661
- Thermodynamics
 - Boltzmann’s and Nernst’s work 2:131
 - FPU phenomenon 2:127
 - for fractal analysis in dynamics 2:396
 - nearly integrable *vs.* hyperbolic
 - systems 2:126
 - nonequilibrium statistical
 - mechanics 3:532
 - overview 2:125–133
 - physical interpretation 1:58
- Thermohydraulics, Newtonian fluids
 - adjoint system 3:497
 - Boussinesq approximations 3:493
 - Boussinesq local balances 3:493
 - density state equations 3:493
 - dimensionless local balances 3:495
 - direct system stability 3:496
 - governing equations 3:492
 - gravitational buoyancy convection 3:498
 - hydrostatic state 3:494
 - hydrostatic state stability 3:497
 - linear stability 3:496
 - nonlinear stability 3:497
 - overview 3:492–502
- thermocapillary convection 3:502
- Theta vacuum, chiral anomaly
 - applications 1:207
- Thirring model, classical mechanics
 - overview 1:20
- Thomas–Fermi theory, and stability of
 - matter 5:13
- Thoma theorem, for harmonic functions on
 - Young graph 4:351
- Thom–Boardman classes, and optic
 - caustics 3:622
- Thom classes
 - character form in K -theory 3:391
 - and equivariant cohomology 2:247
 - in Mathai–Quillen
 - construction 3:391
- Three-body problem
 - classical mechanics overview 1:27
 - general three-body problem 2:577
 - planar three-body problem 3:194
 - restricted three-body problem 5:20
 - spatial three-body problem 3:195
- Three-charge system, microscopic entropy,
 - branes 1:378
- Three-dimensional Chern–Simons theory,
 - with gauge group $U(1)$ 4:495
- Three-dimensional gravity, and
 - Chern–Simons theory
 - BTZ black hole 4:501
 - overview 4:501
- Three-dimensional quantum gravity, spin
 - foams in
 - classical theory 4:646
 - covariant path integral 4:650
 - Hamiltonian formulation 4:647
 - kinematical Hilbert space 4:647
- Three-dimensional topological quantum
 - field theory
 - integral structures 4:121
 - state-sum invariants 4:121
 - and 3-manifold invariants 4:119
- Three-dimensional turbulence, wavelets
 - applications 5:415
- Three-manifolds
 - Casson’s invariant 2:477
 - from Chern–Simons theory 4:500
 - finite-type
 - FTI generality 2:349
 - FTIs for ZHS 2:349
 - Goussarov–Habiro definition 2:349
 - Habiro ring 2:354
 - Le–Murakami–Ohtsuki
 - invariant 2:351
 - mathematical theories 2:348
 - overview 2:348–356
 - quantum invariants and perturbative
 - expansion 2:352
 - Floer theory 2:477
 - Jones–Witten theory 2:478
 - quantum three-manifolds 4:117–122

- Three-particle quantum scattering systems,
 - resolvent equations for 3:588
- Threshold bound state, definition 1:382
- Thresholding, and wavelet
 - denoising 5:436
- Thurston's theory, and postcritically finite maps 2:656
- Tidal effect, and general relativity 2:488
- Tilings, in combinatorics 1:568
- Time asymptotics, and resonances 4:417
- Time delay
 - in gravitational lensing 2:571
 - multiply-imaged quasars 2:573
- Time-dependent Ginzburg–Landau equations, overview 2:551
- Time-dependent scattering theory, and Møller operators 4:252
- Time discretization, and wavelet computation 5:417
- Time-independent scattering theory, overview 4:253
- Time machine question, overview 4:210
- Toda lattices
 - affine Lie algebras 5:241
 - band matrices 5:243
 - classical mechanics overview 1:14
 - finite open-end 5:238
 - finite periodic 5:239
 - infinite 5:236
 - integrable discretization 3:63
 - with nearest-neighbor interactions 5:241
 - overview 5:235–244
 - and Sato's theory 5:242
 - simple Lie algebras 5:240
 - 2D 5:243
- Toda theories, affine Toda theories 3:51
- Toeplitz determinants, and statistical mechanics
 - nonsmooth symbols 5:247
 - nonzero index symbols 5:249
 - overview 5:244–251
 - and strong Szegő limit theorem 5:245
- Tolmann–Oppenheimer–Volkoff metric and FRW metric, overview 4:561–562
- FRW–TOV shock matching 4:565
- Tolmann–Weitsman theorem, and Hamiltonian group actions 2:606
- Tomita–Takesaki modular theory 5:252
 - algebras and states 5:254
 - applications 5:257
 - basic structure 5:251
 - geometric modular action 5:256
 - modular automorphism group 5:252
 - modular invariants 5:253
 - in quantum theory, overview 5:255
 - and self-dual cones 5:254
 - von Neumann algebra
 - classification 5:253
- Topological defects
 - composite defects 5:263
 - examples 5:261
 - formation 5:258
 - helium-3 5:262
 - homotopy groups 5:258
 - overview 5:257–264
- quantum fields
 - Anderson–Higgs–Kibble mechanism 4:225
 - boson transformation theorem 4:224
 - overview 4:221–229
 - singularities and massless bosons 4:224
 - symmetry and order 4:221
 - symmetry rearrangement 4:223
 - two-level theory approach 4:222
 - spontaneous symmetry breaking 5:257
- Topological degree of mapping, and Leray–Schauder theory
 - extensions 3:283
 - overview 3:281–285
- Topological dynamics
 - regular and stochastic dichotomy 5:348
 - and stochastic maps 5:348
 - and unimodal maps 5:343
- Topological field theories, and Batalin–Vilkovisky formalism 1:251
- Topological gauge theory, Mathai–Quillen formalism applications 3:398
- Topological gravity, 2D 5:264–270
- Topological groupoids, examples 3:314
- Topological invariants
 - and Hamiltonian system integrability 2:628
 - subfactors 5:388
- Topological knot theory
 - and fluid flows 5:274
 - and fluid mechanics 5:274
 - in fluid patterns 5:277
 - invariants and vector fields 5:272
 - and magnetic flux 5:274
 - magnetic knot and link energy 5:276
 - overview 5:271–278
 - vortex dynamics ideas 5:271
- Topologically chaotic unimodal maps, characteristics 5:343–344
- Topological quantum field theory
 - axiomatic approach 1:232–234
 - Chern–Simons gauge theory 5:284
 - and Donaldson–Witten theory 5:282
 - duality overview 2:118–125
 - extended supersymmetry twisting 5:281
 - and Jones polynomial 3:185
 - and knot theory 3:227
 - in low dimensions 1:233
 - and modular functor 1:359
 - overview 1:352, 2:390, 5:278–290
 - Schwarz-type topological QFT 4:494–503
 - Seiberg–Witten invariants 5:287–288
 - sigma models 5:286
 - string theory 5:286–287
 - structure and computations 5:279
 - structures 1:233
 - 3D QFT 4:119
 - twisted theories 2:119
 - and Wilson loops 5:288–289
- Topological reasoning, in general relativity 4:617
- Topological sigma models
 - A and B models 5:295
 - construction 5:291
 - Mathai–Quillen formalism applications 3:397
 - observable construction 5:292
 - observables and intersection theory 5:293
 - overview 5:290–295
 - as TQFT type 5:286
- Topological space
 - functional analysis overview 1:89
 - and Lie groupoids 3:314
- Topological sphere theorem, definition 3:347
- Topological strings
 - and large-N duality 3:266
 - and large-N expansion, overview 3:263–268
 - theory example 5:286–287
- Topological structures, for manifolds 2:386
- Topological theory
 - knot group roles 3:400
 - knots and links 3:399
- Topological vortex dynamics, basic ideas 5:271
- Topology concepts
 - algebras 1:139
 - compact complex manifolds 4:557
 - essential concepts 1:131
 - fields 1:139
 - in fluid flows 5:274
 - function spaces 1:135
 - graded Poisson algebra 2:564
 - groups 1:138
 - homotopy groups 1:135
 - instanton topology 3:44–50
 - and integrable systems 2:432
 - intersection theory links 3:153
 - lattice gauge theory 3:280
 - low-dimensional, instantons and 4-manifolds 2:475
 - maps 1:138
 - metric spaces 1:133
 - modules 1:139
 - moduli spaces 2:474
 - operator spaces 1:135
 - overview 1:131–139
 - relations 1:138
 - rings 1:139
 - sets 1:138

- spacetime topology 4:617–623
- string topology 5:111–114
- and symmetries 5:170
- vector spaces 1:139
- Toric manifolds
 - and Delzant's theorem 2:604
 - elementary properties 2:604
- Tori
 - Kolmogorov–Arnol'd–Moser tori 3:584
 - lower-dimensional tori 5:28
 - maximal KAM theory tori 3:191
 - maximal tori 1:578
 - noncommutative tori 3:524–530
- Torsion, and Einstein–Cartan theory 2:191, 2:194
- Torus knots, knot theory 3:401
- Total mass, in general relativity 2:510
- TOV metric *see*
 - Tolmann–Oppenheimer–Volkoff metric
- Tower constructions, and ergodic theory 2:252
- TQFT *see* Topological quantum field theory
- T – Q relation, overview 2:161
- Trace anomalies, characteristics 1:209
- Trace-class method, quantum mechanical scattering theory 4:254
- Trace formulas
 - and resonances 4:418
 - Selberg trace formula 1:214, 4:513
 - semi-classical 4:512–518
- Trace manifolds, and Feynman integral 4:516
- Traces, in space BV 2:412–413
- Transcritical bifurcation, and singularity theory 4:590
- Transfer matrix method, in combinatorics 1:564
- Transformation formulas
 - Bailey's ${}_{10}W_9$, elliptic analog 4:115
 - in q -gauss hypergeometric series 4:108
 - for ${}_r\phi_{r-1}$ series 4:108
- Transformation groups
 - differential geometry overview 1:35
 - for soliton equations 4:598
- Transformations
 - and algebraic QFTs 1:203
 - Austenite–Martensite transformations 5:363
 - Bäcklund transformations 1:241–247
 - boson transformation 4:224
 - coarse graining transformations 4:55
 - combescure transformations 3:82
 - Darboux-type transformations 3:82
 - and Einstein–Cartan theory of gravitation 2:192
 - in Fedosov quantization 2:295
 - for finite periodic Toda lattice 5:240
 - gauge transformations 1:50, 2:468, 4:242
 - induced transformations 2:252
 - in integrable dynamical systems 3:107–108
 - Laplace transformations 3:83
 - Lévy transformations 3:82
 - Miura transformations 4:3
 - phase space coordinates 1:4
 - solid–solid phase transformations 5:368
- Transition chain, and Hamiltonian system instability 2:635
- Transition cocycles, vector bundles 1:488
- Transitions
 - and large- N dualities 3:267, 3:272
 - liquid–vapor transition 4:56
 - phase transition *see* Phase transition dynamics
- Transitivity
 - and hyperbolic dynamical systems 2:723
 - and stable ergodicity 2:727
- Translation invariance, and quantum spin system symmetry 4:298–299
- Translations, and quantum n -body problem 4:284
- Transmission systems, chaos
 - synchronization applications 5:225
- Transonic shock, inviscid compressible fluid flows 1:600
- Transport equation, in fluid mechanics 2:371
- Transpose, partial, positivity 2:230
- Transversality *see* Quantitative transversality
- Trapped surfaces
 - and general relativity 2:505
 - and spacetime topology 4:619
- Trappons, and integrable PDEs 3:116
- Tree level
 - definition 1:382
 - in perturbation theory 2:56
- Triangularity, quantum Calogero–Moser systems 4:126
- Tricomi's equations, and symmetric–hyperbolic systems 5:165
- Triviality, singular reduction 5:196
- TSG effect *see* Tsui–Stormer–Gossard effect
- Tsui–Stormer–Gossard effect
 - in higher Landau levels 2:410
 - overview 2:404
- Tubes, magnetic flux 5:274
- Turbulence
 - bidimensional turbulence 5:77
 - definition 5:408
 - energy method 5:351
 - intermittency in, overview 3:144–151
 - internal intermittency 3:144
 - and inviscid flows 3:165
 - and knots 5:274
 - methods of study 5:408
 - turbulent momentum transport 5:352
 - variational methods
 - overview 5:351–355
 - wavelets applications
 - coherent structure extraction 5:414
 - compression principle 5:413
 - and Fourier spectrum 5:411
 - intermittency measures 5:413
 - overview 5:408–420
 - spectra 5:410
 - and structure functions 5:412
 - 3D turbulence 5:415
 - 2D turbulence 5:418
 - wavelet computation 5:417
- Turbulence theories
 - closure problem 5:298
 - equations of motion 5:297
 - intermittency 5:302
 - Kolmogorov theory 5:300
 - mathematical aspects 5:302
 - overview 5:295–303
 - Reynolds equations 5:298
 - Reynolds number 5:297
 - statistical theory elements 5:299
 - turbulent regimes 5:296
- Turbulent flows, properties 5:296
- Turbulent momentum transport
 - upper bounds 5:353
 - variational problem 5:352
- Turing machines, cellular automata as 1:464
- Twisted K -theory
 - and Chern character 3:251
 - and string topology 5:114
 - and type II string theory duality 3:253
- Twisted signature elliptic complex, characteristics 3:25
- Twisted theories, duality in 2:119, 2:120
- Twistors
 - geometry 5:311
 - linear Penrose transform 5:313
 - nonlinear Penrose transform 5:314
 - overview 5:311–316
- Twistor space, scattering amplitudes in 4:40
- Twistor-string theory, overview 5:308
- Twistor theory
 - completely integrable systems 5:306
 - geometry applications 5:307
 - graviton construction 5:305
 - Klein correspondence 5:304
 - nonlinear equations overview 5:305
 - overview 5:303–311
 - and Penrose transform 5:304
 - twistor-string theory 5:308
- Two-body processes
 - classical gravitational many-body problems 2:576

- Two-body processes (*continued*)
 and massive field theory 4:480
- Two-charge systems, microscopic entropy, branes 1:377
- Two-dimensional conformal geometry
 and chiral fields 5:335
 chiral fields and vertex operator algebras 5:319
 conformal field theory structure 5:318
 conformal symmetry group 5:317
 fusion rules and Verlinde's formula 5:321
 and modular invariance 5:321
 overview 5:317–322
 rational theories 5:320
 Zhu's algebra 5:320
- Two-dimensional Euler equations, weak solutions 3:14
- Two-dimensional inviscid flows, Arnold's work 3:163
- Two-dimensional Ising model
 critical-exponent phenomenology 5:325
 and Fuchsian equations 5:326
 holonomic quantum fields 2:660
 overview 5:322–328
 and scaling theory 5:324
 solvability 5:322
 susceptibility 5:325
- Two-dimensional models
 boson–fermion equivalence and superselection theory 5:331
 chiral theories and conformal quantum fields 5:334
 even theory two-particle bound states 4:485
 integrability 5:338
 manifestation 5:329
 overview 5:328–342
 recent developments 5:339
- Two-dimensional quantum field theory
 bootstrap form-factor construction 5:338
 general concepts 5:329
 overview 5:328
- Two-dimensional Schrödinger operators, and Novikov–Veselov hierarchy 2:37
- Two-dimensional space, percolation duality, exact computations, RSW theory 4:24
 scaling limit 4:24
- Two-dimensional spacetimes, characteristics 3:348
- Two-dimensional Toda lattices, characteristics 5:243
- Two-dimensional topological gravity
 integrable hierarchy 5:268
 intersection theory 5:265
 minimal models 5:267
 overview 5:264–270
 topological σ -model 5:269
 Virasoro conditions 5:269
- Two-dimensional turbulence, wavelets applications 5:418
- Two-particle bound states, for even theories 4:485
- Two-particle Lagrangian dispersion
 in nonsmooth flows 3:258
 in smooth flows 3:256
- Two-step groups, and pseudo-Riemannian nilpotent Lie groups 4:96
- Two-torus
 large complex structure limit point 3:445
 and mirror symmetry 3:444
- ## U
- Uhlenbeck's theorem, and gauge theory 2:469
- Ultraviolet cutoff removal, and renormalization groups 2:276
- Ultraviolet limit, in constructive QFT 1:629
- Ultraviolet stability, in constructive QFT 1:627
- Ultraviolet structure, algebraic QFT 1:201
- $U(N)$, definition 1:382
- Unbounded opens
 exterior domain 2:225
 Sobolev spaces 2:219
 whole space 2:225
- Unbounded operators, and spectral theory 4:639
- Uncertainty principle, and Weyl term 2:703
- Unfolding billiard trajectories, characteristics 4:84
- Uniaxial nematic phase, liquid crystals defects 3:326
- Unidirectional identical synchronization, characteristics 5:217
- Uniformization, Riemann surfaces 4:420
- Uniformly hyperbolic dynamical systems, definitions 2:722
- Uniform motion, field of charges 1:46
- Unimodal maps
 and kneading theory 5:344
 and universality and renormalization 5:343
- Uniqueness
 capillary surfaces 1:436
 central manifolds 1:467
 Einstein manifolds 2:184
 stochastic differential equations 5:67
- Unirreps, classification 1:584
- Unitarity
 and QFT analytic program 4:473
 quantum three-manifold invariants 4:120
- Unitary connections, and characteristic classes 1:493
- Unitary evolutions
 adiabatic evolution 2:530
 Aharonov–Anandan phase 2:532
 Berry's phase 2:531
- Universal calculus, quantum group principal bundles 4:240
- Universal constructors, cellular automata as 1:464
- Universal differential calculus, on algebra 4:237
- Universal formalism, and Calogero–Moser systems 4:125
- Universal geometry
 doubling bifurcations 5:344
 feigenbaum attractor 5:344
- Universality
 classes 5:348
 and Clifford algebras 1:523
 and complex renormalization 5:346
 and Feigenbaum–Coullet–Tresser renormalization conjecture 5:345
 and full renormalization horseshoe 5:345
 and gravitational collapse 1:672
 Le–Murakami–Ohtsuki invariant 2:352
 in random matrix theory 3:103, 4:341
 renormalization mechanism, overview 5:343–349
- Universal mixed-state cloning, quantum states 3:631
- Universal not, and quantum state cloning 3:632
- Universal pure-state cloning, quantum states 3:630
- Universal Vassiliev invariant
 Kontsevich integral 3:234
 universality 3:235
- Universal weight map, finite-type invariants of ZHS 2:350
- UN phase *see* Uniaxial nematic phase
- Unruh effects
 in axiomatic QFT 1:238
 and curved spacetime 4:209
- Unstable fixed points, and renormalization groups 2:277
- Unstable manifolds, and hyperbolic dynamical systems 2:723
- u -plane integral, in Donaldson–Witten theory 2:113
- Upwind flux, in fluid mechanics 2:371–372

V

- Vacuum, dynamical property 1:238
- Vacuum expectation values, and
 - Wightman function definition 4:478
- Vacuum string field theory,
 - overview 5:102
- Vacuum vector, in axiomatic QFT 1:236
- Vafa–Witten theory, as Donaldson–Witten
 - theory extension 2:117
- Valence bond solid states,
 - construction 2:335
- Van der Waals theory
 - equilibrium statistical mechanics
 - overview 1:63
 - and liquid–vapor transition 4:56
- Van Hove’s theorem, equilibrium statistical
 - mechanics overview 1:64
- Variable coefficients,
 - symmetric–hyperbolic system wave
 - equation 5:164
- Variational bounds, spin glasses 4:661
- Variational methods
 - for Ginzburg–Landau energies
 - functional with magnetic field 5:360
 - issues and methods 5:357
 - overview 5:355–362
 - for microstructures
 - Austenite–Martensite
 - transformations 5:363
 - calculus of variations 5:364
 - minimizers 5:366
 - overview 5:363–368
 - relaxation 5:365
 - in turbulence
 - energy method 5:351
 - overview 5:351–355
 - turbulent momentum transport 5:352
- Vassiliev invariants
 - and knot theory 3:404
 - and rigid vertex graph invariants 3:223
 - and Witten’s functional integral 3:225
- VBS *see* Valence bond solid states
- Vector bundles
 - and characteristic classes 1:488
 - Chern classes 1:492
 - cohomologies 1:491
 - direct sum and tensor product 1:490
 - frame bundles 1:490
 - frames 1:488
 - for heterotic strings 2:382
 - homotopy 1:490
 - K -theory 1:490
 - Lie groupoids 3:314
 - line bundles 1:492
 - orientability 1:492
 - Pontrjagin classes 1:492
 - principal and associated bundles 1:489
 - principal bundles examples 1:488
 - projective space tangent bundle 1:489
 - pullback bundle 1:490
 - simple covers 1:488
 - space classification 1:491
 - spaces with finite covering
 - dimension 1:491
 - spin structures 1:489, 1:492
 - stabilization 1:490
 - Stiefel–Whitney classes 1:491
 - structure group lifting 1:489
 - structure group reduction 1:489
 - symplectic reduction 3:456
 - transition cocycles 1:488
- Vector cellular automata, overview 1:458
- Vector coherent states,
 - equations 1:543–544
- Vector fields
 - division algebras 1:528
 - and geometric measure theory 2:523
 - and knot invariants 5:272
 - and singularity theory 4:590
 - spheres 1:528
 - symmetry reduction 5:191
- Vector nonlinear Schrödinger systems,
 - characteristics 3:558
- Vector space
 - and Clifford algebras 1:523
 - in topologies 1:139
- Vector states, C^* -algebras 4:90
- Verlinde formula
 - and conformal field theory 5:321
 - and Riemann surfaces 4:428
 - in semisimple abelian modular
 - categories 1:357
 - and string topology 5:114
- Versal deformations, definitions and
 - examples 2:20
- Vertex models, and Yang–Baxter
 - equations 5:467
- Vertex operator algebras
 - and conformal field theory 5:319
 - overview 5:317–322
- VEVs *see* Vacuum expectation values
- Vibrating string, partial differential
 - equations control 1:640
- Vicious walkers, in combinatorics 1:570
- Violent relaxation, classical gravitational
 - many-body problems 2:580
- Virasoro algebra
 - and bi-Hamiltonian
 - formulation 1:292–294
 - in conformal field theory 1:334
 - overview 1:676
 - and 2D topological gravity 5:269
- Virial theorem
 - classical gravitational many-body
 - problems 2:578
 - equilibrium statistical mechanics
 - overview 1:62
- Viscoelastic flows, theoretical and
 - numerical problems 3:566
- Viscosity fluids, and constitutive
 - equations 3:563
- Viscosity models, in vortex
 - dynamics 5:398
- Viscous compressible fluid flows,
 - Navier–Stokes equations,
 - overview 1:602
- Viscous incompressible fluids
 - basic problem 5:369
 - and continuous data dependence
 - 5:370
 - and Galerkin approximation 5:376
 - overview 5:369–379
 - pointwise in time estimates 5:372
 - regularity 5:373
 - and Reynolds number 5:370
 - smoothing estimates 5:373
- Vlasov equations, and mean-field
 - scaling 3:203
- Vlasov–Poisson system, classical
 - gravitational many-body
 - problems 2:579
- VNLS systems *see* Vector nonlinear
 - Schrödinger systems
- VOF method *see* Volume-of-fluid method
- Volterra maps, and functional
 - integration 2:437
- Volume
 - and Clifford algebras 1:524
 - and noncommutative geometry 4:9
- Volume conjecture, for finite-type
 - invariants 2:347
- Volume constraint, capillary
 - surfaces 1:434–435
- Volume elements, and functional
 - integration 2:438
- Volume-of-fluid method, for
 - multicomponent fluid flow 3:136
- Volume-preserving diffeomorphisms, as
 - infinite-dimensional Lie group
 - example 3:42
- Von Klitzing effect, overview 2:403
- Von Neumann algebras
 - abstract types 5:382
 - and cellular automata 1:464
 - Connes’ classification 5:383
 - injective factors 5:383
 - and modular theory 5:383
 - Murray–von Neumann
 - classification 5:380
 - overview 5:379–384
 - positive semigroups on 4:93
 - and Tomita–Takesaki modular
 - theory 5:253, 5:254
- Von Neumann algebra subfactor theory
 - algebraic QFT 5:388
 - analytic classification theory 5:387
 - constructions and combinatorial
 - classification 5:387
 - overview 5:385–390

- Von Neumann algebra subfactor theory
(*continued*)
topological invariants 5:388
- Von Neumann entropy,
characteristics 4:178
- Vortex dynamics
aircraft trailing vortices 5:393
associated statistics 5:77
atmospheric vortex 5:395
background 5:390
basic ideas 5:271
boundary point motion 4:73
contour dynamics in 2D 5:397
filament methods in 3D 5:398
and Ginzburg–Landau
model 2:550, 5:357
Higgs vortices 1:151–160, 4:605
and mixing and chaos 5:394
N point vortices 4:73
numerical methods 5:396
overview 5:390–399
point-vortex dynamics 4:66–79, 5:396
and quantum fields 4:225
shear layers 5:392
and simplified Ginzburg–Landau
model 5:357
stretching in turbulent flows 5:296
superfluids and superconductors 5:395
topological vortex dynamics 5:271
viscous vortex methods 5:398
vortex rings 5:394
- Vortex lines, and knots 5:274
- Vortex rings
characteristics 5:394
in point-vortex dynamics 4:71
- Vortex scattering, abelian Higgs
vortices 1:155
- W**
- Wall-crossing formula, and u -plane
integral 2:115
- Ward correspondence
and gauge theory 2:472
as integrable system 2:473
- Ward identities
classical mechanics overview 1:17
in conformal field theory 1:333
operator product expansion 3:619
- Wasserstein distance, and optimal
transportation 3:636
- Water waves
breaking water waves 1:383–386
dissipative water-wave models 2:136
as dynamical systems 2:133–138
Hamiltonian systems in
hydrodynamics 2:135
hydrodynamics asymptotical
models 2:133
initial-value problem 2:134
nonlinear Schrödinger equations
in 3:554
shallow-water waves 4:13
stochastic water-wave models 2:137
- Watson’s transformation, terminating ${}_8W_7$
into terminating balanced ${}_4\phi_3$ 4:109
- Wave collapse, and nonlinear Schrödinger
equations 3:558
- Wave coordinatization, equations 1:343
- Wave equations
diffraction, geometrical optics 5:404
and diffraction, overview 5:401–407
and Hamilton’s equations 3:38
and linear evolution
equations 2:266–267
relativistic wave equations 4:391–398
semilinear wave equations 4:518–526
and symmetric-hyperbolic
systems 5:164
- Wave fronts
and hyperbolic billiards 2:718
oblique collisions, and peakons 4:19
perfect fluid 2:197
singularity locus, caustics as 3:623
- Wave functions
quantum Calogero–Moser systems
4:126
in random matrix theory 4:341
- Wave group, and geodesic flow 4:184
- Wavelet applications
multidimensional CWT 5:422
multifractal function analysis 5:421
multiscale analysis 5:422
and numerical calculations 5:425
overview 5:420–426
pattern detection 5:423
signal compression 5:424
singularity detection 5:421
to turbulence
coherent structure extraction 5:414
compression principle 5:413
and Fourier spectrum 5:411
intermittency measures 5:413
overview 5:408–420
spectra 5:410
and structure functions 5:412
3D turbulence 5:415
2D turbulence 5:418
wavelet computation 5:417
- Wavelets
coefficients in regularity and local
decay 5:434
continuous wavelet transform 5:426
decomposition, and image
processing 3:7
definition 5:409
and denoising 3:9, 5:436
discrete wavelets 5:429
in image processing 3:7
and linear approximation 5:434
mathematical overview 5:426–438
and nonlinear approximation 5:434
operator compression and
preconditioning 5:435
polynomial reproduction 5:433
types and generalizations 5:425
- Wavelet space, and turbulence 5:410
- Wave–particle dualism, in Hubbard
model 2:713
- Wave propagation
boundary control method 1:340–345,
1:345
forward problems
controllability 1:341
dynamical system 1:340
Laplacian on waves 1:341
manifold 1:340
spectral problem 1:342
inverse problems
amplitude formula 1:344
general setup 1:342
model 1:342
solutions 1:344
wave coordinatization 1:343
- WDVV equations, and Frobenius
manifolds
canonical coordinates 5:441
deformed flat connection 5:439
examples 5:444
and integrable systems 5:446
intersection form 5:440
moduli 5:442
overview 5:438–447
- WDW theory *see* Wheeler–De Witt theory
- Weak additivity, and entanglement
measures 2:235
- Weak bisolution property, and $*$ -algebra
construction 4:204
- Weak-coupling limit, and quantum
dynamical semigroups 4:162
- Weak field, classical Lorentz electron
gas 3:467
- Weak gravitational lensing,
characteristics 2:575
- Weak interaction, matrix
elements 3:279
- Weak linking, and saddle point
problems 4:454
- Weakly correlated diffusion, and
random-walk model 3:479
- Weakly coupled oscillators
mean field approximations 5:452
overview 5:448–453
phase resetting 5:448
synchronization 5:451
weak coupling 5:450
- Weak measurements, quantum *see*
Quantum weak measurement
- Weierstrass representation, and minimal
submanifolds 3:427

Weight decomposition, compact groups 1:584

Weighted space, Sobolev spaces 2:219

Weight systems, Kontsevich integral 3:233

Weil algebra, and equivariant cohomology 2:244

Weinstein space, symplectic leaves 1:666

Well-posedness, inviscid compressible fluid flows 1:597, 1:598

Wenzel–Kramers–Brillouin approximations, and functional integration 2:442

Wetting energy, capillary surfaces 1:434–435

Wetting phenomena, statistical mechanics 5:59

Weyl asymptotics, and h -pseudodifferential operators 2:702

Weyl character formula, and compact groups 1:585

Weyl expansion, and semi-classical spectra 4:516

Weyl group, overview 1:580

Weyl law, and quantum ergodicity 4:186

Weyl representations, Clifford algebras 1:525

Weyl solutions, Einstein equations 2:169

Weyl systems *see* Finite Weyl systems

Weyl tensor

- and asymptotically flat spacetime 1:224
- in mathematical cosmology 1:654

Wheeler–De Witt theory

- Einstein gravity canonical structure 5:454
- equations and solutions 5:457
- overview 5:453–462

Wheels formula, Kontsevich integral 3:238

White-dwarf binaries, and gravitational waves 2:584–585

White-noise calculus, for Feynman path integrals 2:312

Whitney conditions, and singular reduction 5:196

Whole space

- Ginzburg–Landau equations 2:548
- unbounded open 2:225

Wick quantization, quantum mechanics overview 1:123

Wick’s theorem, in perturbation theory 4:30

Wiener space, differential calculus on 3:384

Wightman axioms

- and dispersion relations 2:96
- in massive field theory 4:478

Wigner’s concept of particle, and scattering theory 4:462

Wild behavior

- C^1 -generic diffeomorphisms 2:501
- surface diffeomorphisms 2:498

Wilson loops

- and AdS–CFT 1:179
- and topological QFT 5:288–289

Wilson polynomials, characteristics 3:644–645

Wilson’s formulation, for lattice gauge theory 3:276

Witten index, in supersymmetric quantum mechanics 5:150

Witten’s functional integral, and Vassiliev invariants 3:225

WKB approximation *see* Wenzel–Kramers–Brillouin approximations

Wrapped Dirichlet branes, D5, and super Yang–Mills 2:466

Wulff droplets

- crystals in Ising model 5:463
- overview 5:462–464
- and percolation theory 4:22

Wu–Yang criterion, in abelian duality 2:204

Y

Yang–Baxter equations

- in boundary quantum groups 1:189
- checkerboard generalizations 5:468
- checkerboard IRF model 5:469
- checkerboard IRF-vertex model 5:469
- checkerboard model square weight 5:470
- checkerboard vertex model 5:469
- checkerboard vertex model map 5:470
- classical YBEs 5:471
- eight-vertex and hard-hexagon models 2:158
- and electric networks 5:465
- and factorizable S -matrices 5:466
- higher-dimensional generalizations 5:472
- and interaction-round-a-face model 5:467
- inverse relations 5:472
- and IRF-vertex model 5:468
- and knot theory 5:466
- operator formulations 5:471
- overview 5:465–473
- and quantum inverse-scattering method 5:468
- reflection YBEs 5:471
- spin model map 5:470
- and spin models 5:465
- and vertex models 5:467

Yangians, in quantum affine algebras 1:186

Yang–Mills complex characteristics 3:25

- index 3:28

Yang–Mills equations

- curved twistor spaces 5:315
- with differential forms 1:148
- and Dirac fields 2:68
- and electromagnetism 1:149
- and gauge theory 2:469
- in twistor theory 5:306
- variational aspects, and instanton topology 3:45

Yang–Mills–Higgs Ansatz, and standard model 3:510

Yang–Mills instantons, characteristics 4:606

Yang–Mills theory

- and electromagnetism 5:167
- in four dimensions 3:570
- and gauge theory in three dimensions 3:576
- instantons 3:529
- Morita equivalence 3:527
- and nonabelian BF theories 1:262
- on noncommutative tori, overview 3:526
- and quantum chromodynamics 4:503
- and Seiberg–Witten duality 4:507

YBEs *see* Yang–Baxter equations

Yield stress, and non-Newtonian behavior 3:561

Yield stress fluids, and Bingham model 3:565

YM equations *see* Yang–Mills equations

Young graphs, harmonic functions 4:351

Y-surgery, and finite-type invariants 2:349

Z

Zero-limit temperature, fermion renormalization 4:412

Zero-mode approximation, random matrix theory supersymmetry method 5:159

Zero-noise limits, and random dynamical systems 4:336

Zero-one laws, random walks in random environments 4:368

Zero-range process, and thermodynamic fluctuations 3:364

Zero-rest-mass free fields, linear Penrose transform 5:313

Zeta values, in number theory 3:601

Ziglin’s theory, and Hamiltonian system integrability 2:629

世

界著名数学家、哲学家、逻辑学家弗雷格曾给出了一个著名等式：半个数学家+半个哲学家=好的哲学家+好的数学家。他解释说：“一个好的数学家，至少是半个哲学家；一个好的哲学家，至少是半个数学家。”

本书的目的就是要用物理学家替换上述等式中的哲学家。

举两个刚刚读到的例子，从中可见物理学家对数学也会有贡献。物理学家李政道和杨振宁在研究统计力学的一个问题时，遇到了一类特殊的多项式

$$P(z) = \sum_{j=0}^m a_j z^j$$

的集合 \mathcal{P} 。他们能够分析出， \mathcal{P} 中的任意一个多项式 P 的所有根都位于复平面的单位圆周 $\{z: |z|=1\}$ 上。因此他们猜测这个结论对 \mathcal{P} 中的所有多项式 P 都成立。如果他们可以找到一个酉矩阵 U 使得 $P(z)$ 是 U 的特征多项式，即 $P(z) = \det(zI - U)$ ，那么猜想就证明了。这是任何一个学过高等数学的人都会想到的办法，但这个方法在此不管用。杨和李有很好的数学功底，因此找到一个证明，但这个证明并不简单。现在有更容易的证明了，这要特别归功于浅野太郎 (Taro Asano)。为证明杨-李单位圆定理 (将在下面陈述)，我们需要将单变量 z 的 m 次多项式 P 替换为 m 个变量 z_1, \dots, z_m 的多项式 $Q(z_1, \dots, z_m)$ ， $Q(z_1, \dots, z_m)$ 关于每个变量 z_i 都是一次的。我们感兴趣的是这样一类多项式 $Q(z_1, \dots, z_m)$ 的集合 \mathcal{Q} ：只要 $|z_1| < 1, \dots, |z_m| < 1$ 就有 $Q(z_1, \dots, z_m) \neq 0$ 。因此，如果 $P(z) = Q(z, \dots, z)$ 且在 \mathcal{Q} 中，则 P 的根 ξ 满足 $|\xi| \geq 1$ 。(在我们感兴趣的情况下，存在一个对称 $z \rightarrow z^{-1}$ ，因此也有 $|\xi^{-1}| > 1$ ，从而 $|\xi| = 1$ 。)很明显，如果 $Q(z_1, \dots, z_m)$ 与 $\tilde{Q}(z_{m+1}, \dots, z_{m+n})$ 在 \mathcal{Q} 中，则

$$Q(z_1, \dots, z_m) \tilde{Q}(z_{m+1}, \dots, z_{m+n})$$

也在 \mathcal{Q} 中。我们现在描述一个不那么显然的运算，称之为浅野缩并，它将 \mathcal{Q} 中的多项式变为 \mathcal{Q} 中的多项式。记

$$Q(z_1, \dots, z_m) = Az_j z_k + Bz_j + Cz_k + D$$

其中 A, B, C, D 是变量 z_1, \dots, z_m 中除去 z_j, z_k 之外的其余 $m-2$ 个变量的多项式。浅野缩并将两个变量 z_j, z_k 替换为一个单独的变量 z_{jk} ，使得

$$Az_j z_k + Bz_j + Cz_k + D \rightarrow Az_{jk} + D$$

从一个 m 元多项式 Q 出发，经过一次浅野缩并，我们得到一个 $m-1$ 元多项式，如果原来的多项式在 \mathcal{Q} 中，则所得的新的多项式也在 \mathcal{Q} 中。(这是一个简单的练习： $Az_{jk} + D$ 的根是 $Az^2 + (B+C)z + D$ 的两根之积的相反数。)可以验证，如果 $-1 \leq a_{jk} \leq 1$ ，则两个变量 z_j, z_k 的形如

$$z_j z_k + a_{jk} (z_j + z_k) + 1$$

的多项式也在 \mathbb{Q} 中。(令多项式等于零,则得到一个映射 $z_j \rightarrow z_k$,该映射是一个对合,并且将单位圆的内部映射到单位圆的外部。)将这些多项式相继相乘,当同一个变量出现两次时做一次浅野缩并,最后令所有的变量都等于 z ,则我们得到杨—李单位圆定理:对于实数 $a_{jk} = a_{kj}$, $-1 \leq a_{jk} \leq 1$, 多项式

$$P(z) = \sum_{X \subset \{1, \dots, m\}} z^{|X|} \prod_{j \in X} \prod_{k \notin X} a_{jk} \tag{*}$$

的所有根都位于单位圆周上^①。

再比如物理学家张宗燧。张宗燧步入量子场论研究领域,主要受到玻尔(N. Bohr)的影响。从两人的通信中,可以看出张宗燧对理论研究的偏好。而在理论研究中,张宗燧又有明显的数学倾向。其研究特点为:数学技巧强,善于应用数学解析物理理论问题。在物理研究中,他主张多做群论和对称性的工作。其研究成果中数学计算和表达都相当“清楚、干脆、可靠”,结论简明准确。在《数学译林》为田方增先生百岁诞辰的贺信中就提到:泛函分析学科在中国科学院数学研究所几乎一开始就是基础理论与应用并重地发展。按科学规划的精神,从1958年起数学所泛函分析学科强调其发展要侧重于与微分方程、物理学、高尖科技和国民经济建设之联系。为此,田方增、关肇直常与吴新谋、张宗燧等合作,使数学所内泛函分析的发展始终注意与微分方程及现代数学物理的联系。先后组织了量子场理论、粒子迁移理论和电磁波理论中数学问题之研究等学术讨论班。他撰写的学术论文为发展中国在这一领域的数学研究做出了重要贡献。田方增与关肇直一起成功地在中国开辟了应用泛函分析的一个重要领域——粒子迁移理论的数学基础及问题的研究。

所以说数学和物理互易性强。一些数学家后来成了物理学家(例如戴森(Freeman Dyson)),而另一些人正好相反(例如钱德拉(Harish Chandra)、博特(Roul Bott)),他们从物理学家变成了数学家。最夸张的莫过于威腾(Edward Witten, 1951—), 1990年获得菲尔兹奖的理论物理学家威腾于1976年在普林斯顿大学在诺贝尔奖得主(2004)格罗斯(David Gross)的指导下获得物理学博士学位;但他从未获得过数学博士学位。

那么学习物理到底应该掌握多少数学呢?

一位致力于学习理论物理的学生曾请教赫柏林院士怎样治学。赫先生说:“要想搞理论物理,首先数学要好。前两年先把斯米尔诺夫的五卷及变分学、微分几何、数理方法、拓扑和积分等学完,然后开始进入近代数学,要学流形、群、连续群、李群、现代微分几何等。”

当然这只是入门级的数学。

本套丛书貌似物理实则充斥着现代数学,正如中国科学院理论物理研究所吴岳良研究员所评介的那样:

本书物理学部分与数学部分的关系很难分开。实际上,经典力学、电磁学、统计力学、量子力学、流体力学、可积系统和动力系统许多物理问题可归结为求解数学上的常微分方程、偏微分方程、积分方程、微分积分方程等数学物理方程,物理学问题的解会涉及复变函数和特殊函数等多种函数,在求解时又会用到变分技术、调和分析、泛函分析等各种数学分析方法。同时,对爱因斯坦狭义相对论和广义相对论,它不仅改变了人们的时空观,还使得闵可夫斯基时空的几何学和黎曼空间的几何学成为物理理论的数学基础,同时也使得向量分析、张量分析和微分几何等成为必要的数学分析工具。在量子力学中,物理量成为算子,物理状态用波函数来描述,算子的谱才是测量到的物理量。在量子场论中,波函数又被二次量子化成为算子用来描述基本粒子在相互作用过程中的产生和湮灭。这使得算子代数、量子化方法和路径积分等数学理论和方法成为量子物理的数学基础。粒子物理学家发现自然界的3种基本作用力:电磁相互作用、弱相互作用和强相互作用可用规范理论来描述,并完全由规范对称性来支配,这些对称性在数学上用李群和李代数来描写。事实上,晶体的结构也是由欧几里得空间中的转动群来描述,这使得群论在物理学中的应用,尤其在粒子物理中的应用变得越来越重要。在规范理论中,规范势当作基本的量子场,而它被发现就是数学家在现代微分几何学中所研究的纤维丛上的联络,这使得有关纤维丛的拓扑不变量在粒子物理和量子场论研究中变得重要起来,如规范场的磁单极子和瞬子解及手征量子反常等。在量子引力和超弦理论的研究中,不仅运用到已有的数学理论和方法,尤其是现代数学,还促进了数学理论本身的发展。同样,在凝聚态物质和光学方面,物质的拓扑相和拓扑缺陷、拓扑量子计算等也应用到了许多现代数学方法,这使得代数拓扑、代数方法、量子群、复几何、辛几何与拓扑、低维几何、非交换几何等数学理论和数学方法越来越多地渗透到理论物理的研究中。另外,在研究微观物理对象的随机性和各种随机过程的统计规律、无序系统和动力系统时,随机方法和离散数学等也得到越来越广泛

① 见杨振宁,李政道,“Statistical Theory of Equations of State and Phase transition. II. Lattice Gas and Ising Model”, Phys. Rev. (2) 87(1952),410-419;也见 T. Asano, “Theorems on the partition functions of the Heisenberg ferromagnets”, J. Phys. Soc. Japan, 29 (1970),350-359. 长期以来我都为杨—李单位圆定理着迷(见 D. Ruelle, “Extension of the Lee-Yang circle theorem”, Phys. Rev. Lett. 26 (1971),303-304),而且我认为在这个领域仍然有未被揭示出的神秘。(2010年,吕埃勒再次发表了一篇关于杨—李单位圆定理的文章,见 Characterization of Lee-Yang polynomials, Annals of Mathematics, 171(2010),589-603. ——译者注。)

的应用。

数学对物理的影响有多大？

正如本书前言中所写：

当然，数学是确实存在的。事实上，从某种角度而言，物理学是由精确的数学逻辑所操控的：古希腊人把空间几何结构变成了一种真实的艺术形式。就我所知，古希腊人是“数学物理”的第一个践行者，他们引入了坐标轴的概念，从而把空间几何的所有量都转化为一些简单的数字。今天，这些被称作“物理学的基本定律”，直到很久以后我们才认识到如下事实：时间流可以类似地被坐标化，它连同空间一起，同样可用几何方法来解决。于是，有一些疯狂的人对数字的魔力很感兴趣，但是，我们的现实世界似乎确实包含许多超出我们分析能力的地方。

渐渐地，所有这一切都变了。月亮和其他行星的运动好像都满足几何定律。伽利略和牛顿设法去发现这些运动的合乎逻辑的定律，并注意到质量的概念也适用于太空中的物体，就像地球上的苹果和大炮一样，这使得太空更容易被我们所理解。同时人们发现，电子、磁场、光和声音也完全按照数学方程在运转。

科学家认为：开展对“数学物理”的深入研究，有助于揭示出物理学与数学之间的内在联系。事实上，从自然哲学发展到物理学，除了使用实验手段和新的思维方法，数学起了不可替代的作用。当人们通过分析大量实验数据和吸取各种唯象理论的精髓，以严格的数学语言和简洁的数学公式描述支配物质基本结构和宇宙演化的物理规律时，物理学的简洁美、统一美、对称与不对称美则通过深刻的数学美反映出来。可以说，自从物理学成为自然科学的一门独立学科后，物理学与数学之间的关系变得密不可分。古代的许多科学家既是数学家也是物理学家，尤其到了近代和现代，许多理论物理学家对数学的运用和发展起到了更为积极的推进作用，数学家和理论物理学家之间的合作也变得越来越频繁、越来越深入，他们成为了“数学物理”的践行者。大家最为熟知的古希腊的阿基米德，他既是著名的数学家也是著名的物理学家，他很早就利用数学这个工具证明了杠杆原理和浮力原理，并做了大量的实验。牛顿在研究物体和天体的运动规律时发展出新的数学方法——微积分。爱因斯坦则运用对当时的物理学家来说全新的数学方法——微分几何和黎曼几何，创立了广义相对论。爱因斯坦曾回忆说：“1912年我突然认识到，高斯的曲面理论是解开这个秘密的钥匙，他的曲面坐标系意义重大。不过，当时我还不知道黎曼已经更深入地研究了几何基础。我突然想起，读大学时盖泽先生给我们上的几何就包括高斯理论……我认识到几何基础具有物理学意义。当我从布拉格回到苏黎世时，我亲爱的朋友、数学家格罗斯曼也在苏黎世。他告诉了我高斯，然后是黎曼。格罗斯曼两肋插刀，直接催生广义相对论。”

伟大的几何学家海曼·格拉斯曼^①在1844年发表的《Lineale Ausdehnungslehre》(《延拓理论》)，这本书像麦比乌斯的那本名著一样具有丰富的思想，但与麦的写作风格不同，非常晦涩，以至几十年未被人注意，也没有被人读懂，只是在其他书和文章中出现了一系列类似的思想之后，才认识到这些思想出自格拉斯曼的书，不过为时已晚。如果你想领略一下这种抽象的笔法，你只要看一下这本书里的某几章的标题，如：“纯数学之概念之导出”“延拓理论之推导”“延拓理论之叙述”“表示之形式”“一般形式理论之概述”。你只有费劲地钻通了这些内容之后才接触到所述内容的纯抽象的表示，不过仍然很难读懂。直到1862年该书出版了后期的修订本^②，格拉斯曼才用了一种比较容易接受的表示法，即坐标表示法。此外，格拉斯曼选了一个词——Ausdehnungslehre(延拓论)，用以暗示他的研究可应用于任意维空间，而几何学对他而言只不过是这个完全抽象的新学科在普通三维空间中的应用。但是他造的这个新词并没有生根，人们现今简称为“ n 维几何学”。

我们普通读者可能易将数学物理与数学物理方程相混淆，其实这是两个内涵和外延都不同的概念，后者只能视为前者的一个真子集，而前者不论从内容上还是所涵盖的范围都远远超过了后者，但有一点共同之处是它们的问题都源自于物理，但解决都来自于数学家。比如迪利克雷猜想的解决，“迪利克雷原理”这一数学猜想自提出之日起，历经了三十多年的激烈论争和反复，最终才被确立，这是迪利克雷在研究微分方程位势原理时提出的一个猜想，其具体内容简单地大体是：极小化迪利克雷积分

$$\iint \left\{ \left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right\} dx dy$$

的函数 u ，满足位势方程

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

① 海曼·格拉斯曼，《延拓理论》出版于1844年莱比锡，并可参阅其 Gesammelte mathematische und physikalische Werke，第1卷，莱比锡，1894年，第二版出版于1898年莱比锡。

② 柏林，1862年，见其著作集第1卷第二部分，莱比锡，1896年。

后来有人在研究三维位势方程(亦称拉普拉斯方程或调和方程)

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$$

时,又提出,由位势方程所描述的相应物理状态总有一个确定的物理解,因而其本身也必然存在一个数学解,但在数学上的这种存在性,长时间的不能被证明,直到 1851 年,黎曼才在他的博士论文“单复变函数一般理论的基础”中,给出了位势方程边界问题解的存在性证明.由于黎曼在文中运用了他的老师迪利克雷所提出的上述猜想,故他称之为“迪利克雷原理”.可是,在其论文发表后的不长时间,这个原理便激起了热烈的讨论,特别是黎曼的这一证明受到了德国著名数学家魏尔斯特拉斯(K. W. Weierstrass, 1815—1897)的尖锐批评,他指出:黎曼不加证明就先验地假定一定会存在一个使积分取到极小值的函数,这在数学上是不允许的,尽管受到了大师的批评,黎曼并没有因此动摇自己对迪利克雷原理的信心,并且一鼓作气又运用此原理作出了一系列重要的发现.1866 年,黎曼英年早逝,但关于迪利克雷原理是否成立的争论仍未停止.1870 年,魏尔斯特拉斯给出了一个与迪利克雷原理相反的例子,在这个例子中,对给定的边界条件,使迪利克雷积分达到极小值的函数是不存在的,并以此来否定迪利克雷原理.由于迪利克雷原理被当时的数学权威魏尔斯特拉斯所否定,所以数学家们只好另辟蹊径来证明位势方程边界问题解的存在性,比较著名的有三种证法,1870 年纽曼用“算术平均值法”给出了一个证明;1890 年,许瓦兹用“交替法”又给出了一个证明,同年,庞加莱用“扫散法”也给出了一个证明.这些证明从逻辑上讲无疑都是对的,但就是没有一个能够像以迪利克雷原理为工具那样简单、明快,这又不禁使得数学家们怀念起“迪利克雷原理”来,都对它当年被否定而感到惋惜,并随之产生了复活这一原理的念头,并且也为之做出了一些努力,只可惜都未能成功,数学界为此弥漫着一一种悲观的气氛,数学家纽曼就表示:如此优美而又有如此广阔应用前景的迪利克雷原理,已经从我们的视线中“永远消失”掉了!

俗话说“三十年河东,三十年河西”,就在迪利克雷原理被否定三十年之后,即 1899 年,德国领袖数学家希尔伯特对此又发动了一场新的“救亡运动”.他彻底冲破了那种把严格性与简单性对立起来的传统观念,批判了魏尔斯特拉斯以严格性全盘否定迪利克雷原理的做法,从迪利克雷原理的简单性、优美性以及应用的有效性出发,积极寻求它的真实性和合理性,最后终于找到了证明迪利克雷原理的途径和方法,他在德国数学联合会上报了他的这一研究成果,并明确指出:只要对问题中的区域、边界值和允许函数的性质作适当的限制,就完全可以恢复迪利克雷原理的真实性.他还针对数学家们认为迪利克雷原理早已沉没了的观点,意味深长地将他的这一研究工作称为“迪利克雷原理的复活”.后来希尔伯特又给出一个更为一般的证明,从而进一步肯定了迪利克雷原理存在的合理性.

及至近代更多源自于物理的数学理论被抽象出来,而对这些数学理论的进一步研究又极大地推动了物理学的进展,如 Yang-Mills 规范场的大范围整体性质和手征量子反常与纤维丛的拓扑不变量和 Chern-Simons 示性类及指标定理之间建立起直接的联系,超弦理论中的额外维空间与 Calabi-Yan 空间之间的对应关系.理论物理学家威腾在发展超弦理论的同时由于对数学的杰出贡献而获得菲尔兹奖,这些都是物理学与数学相互结合所呈现在“数学物理”方面的经典例子.

对此我国数学工作者早有清醒的认识,20 世纪 80 年代李大潜就撰文指出,学数学的追求纯而又纯的境界,即使从纯数学的发展来说,也不见得是一条康庄大道.不重视实际的需要和其他领域的发展,没有广阔的视野,是很难出第一流的基础理论人才的.

基础和应用有着密切的关系,而且相互促进.搞基础理论的人重视应用方面的教育和训练,对基础理论和应用的研究会带来很大的促进.物理学中的规范场和数学上的纤维丛概念有密切的联系.据杨振宁教授自己讲,他在美国请教了很多纤维丛方面的数学家,但他们讲的一套,他听不懂,双方始终谈不到一起去.只有到了复旦大学,听谷超豪教授用物理学家可以接受的语言,把这二者的关系讲得很清楚,杨振宁教授很高兴,并和谷超豪教授合作,在规范场的数学理论方面做出很多成绩,把这方面的理论进一步发展了.为什么能这样呢?谷超豪教授在念大学时,就选修了物理系四大力学的课程.作为一个数学家,他不仅在数学上有很高的造诣,而且在物理学方面也有很好的修养.

从本书的目录我们可以看出它包含了相当全面的数学内容.它们分别是:数学物理学导言、经典力学、流体动力学、可积系统、经典场论、共形与拓扑场论、量子场论、广义相对论、量子引力、弦论与 M-理论、凝聚态物质与光学、量子信息与量子计算、量子力学、无序系统、动力系统、平衡态统计力学和非平衡态统计力学、代数技巧、李群和李代数、离散数学、量子群、随机方法、复几何、微分几何、低维几何、非交换几何、代数拓扑、辛几何与拓扑、常微分和偏微分方程、泛函分析和算子代数、量子化方法和路径积分、变分技术.

本书的三位主编在序言中写道:“数学物理把数学和物理学这两大学科的优势集中到一起,它们的关系是共同发展.一方面,它运用数学这一工具把不断增长的精确性和复杂性这些物理概念组织了起来;另一方面,物理学家为数学家提供了灵感的源泉.”同时,也正如诺贝尔物理学奖获得者荷兰 Utrecht 大学 Gerard't Hooft 教授在前言中指出:“物理世界与数学世界之间存在明显的重要区别.物理世界强调事实的‘真相’,无论‘真相’是什么,而数学是纯逻辑和纯推理的世界.在物理

学中,一个理论是否能被接受是由实验来最后决定的,物理学中的方法论也与数学不同。”

一个广大读者所关注的例子是天体物理学家霍金是否完美地解决了黑洞火墙悖论?起码现在还没有定论,只能算是给出了第三种可能的解释而已。尽管人们对于黑洞的具体性质还没有全部了解,但是它作为一种致密天体的存在早已没有争议,而黑洞火墙悖论的中心,仍然在于量子力学与广义相对论的矛盾。量子力学把黑洞的视界定义为一个神秘的、拥有巨大能量的火墙,广义相对论则拒绝承认在宇宙中存在这种神奇的火墙,认为黑洞视界只是一种数学上的存在而已。因此,要想真正解决黑洞火墙悖论,人类需要对自然界有更深刻的理解。霍金自己也承认,要想真正理解物质和信息最终从黑洞中逃脱的原理,最终需要人们把引力和自然界的其他作用力合而为一,这是一个困扰了物理学家们将近一个世纪的难题,至今仍然没有得到解决。作为人类现代文明的两块基石,广义相对论通过优美的数学形式描述宇宙,目前人们认为对它已经有足够深刻的理解,而量子力学则通过一种概率化的形式描述微观世界,它的内涵和基本规律仍然不为人知,就连量子力学的创立者尼尔斯·玻尔也说“没有人理解量子力学”。黑洞火墙悖论是这两种理论在宇宙深处的交锋,而交锋的结果,目前仍然无法预料。

本书在刚引进中国时曾有过一个12卷精装本,以内容划分是一种创新,这种事出版界常有。

中央文献研究室所编《毛泽东年谱(1949—1976)》(中央文献出版社,2014)皇皇6卷,是读者期待已久的一部大书。不贤者识其小,这里只摘抄一点儿关于图书装订的内容。1965年8月14日,毛泽东就印一批马列经典大字本问题指示周扬:“同意用照相放大胶印的办法,但请注意封面不用硬纸;大书(例如《唯物主义与经验批判主义》《反杜林论》)过去例作一卷或两卷,现应分装4卷或8卷,使每卷重量减轻。”印大字本,是因为老同志视力差;封面不用硬纸,就是不要硬精装,因其不方便单手握卷、躺着阅读;较厚的书应该多分几册(其实毛泽东推举的两本书都在500页以下)。总体而言,毛泽东对大字本的这些要求,都是以读者为本位,以方便阅读为目的的。有人说:当今出版界在装订方面,流行大开本、大厚本、无线胶订,以傻、大、黑、粗为尚,这种专门为难读者的精神,实在令人费解。

但笔者认为本书绝对算得上是数学物理中的经典之作,而向经典致敬的方式各有不同,最传统、最有效的就是保持原汁原味。原来我们准备连封面都拷贝原版,后与版权代理协商才改成现在的样子。真正美好的东西都一定是增一分则多,减一分则少,原来就刚刚好,我们为什么要破坏它呢?难道我们真的有自信会使其变得更好吗,佛头著粪与狗尾续貂都会让读者吐槽的。

还有一个原因使我们一定要保持原貌,那就是翻译的巨大工作量,我们哈尔滨工业大学出版社地处北方,远离经济与文化中心,实在是没有能力组织庞大的翻译队伍,耗巨资多年打磨这套丛书。我们待将来实力增强后再购买中文版权来完成这一宿愿。在购买版权时我们也表达了购买数字版权的意向,但被婉拒了,因为英文版的数字出版外方已做得很完善了,不像我们刚起步,而且在碎片化之后还面临着版权保护问题,在辞典出版中这是个顽疾。举个例子:

认不认得这个英文单词 esquivallence? 不认得? 那你可以去查一下新版的《新牛津美语词典》(《New Oxford American Dictionary》),里面会告诉你这个词的意思是:“故意逃避自己的官方责任。19世纪开始出现,或许是源自法文 esquiver,‘躲避,溜走’”。

不过如果你拿起家中案头的其他词典,或者将词输入到各种电子词典中,保证你怎么查都查不到这个词,要是你查到了,那可就有事了。

为什么会这样?因为这个词根本就是《新牛津美语词典》编辑部发明的,不存在的词。什么?词典里竟然有虚构的词?编词典的人怎么可以干这种事?

词典里有虚构的词,不只《新牛津美语词典》,基本上每一本词典里都藏有这种凭空创造的词,放这样的词在词典里,倒不是出于编辑的恶作剧坏心,而是有具体用处的。

这是保护著作权的重要机关。辛辛苦苦编出一本厚重的词典,要如何防止别人贪便宜,把你的词典拿去剪剪贴贴,改头换面就变出他们的词典呢?词是共通的,词的意思解释也不会有多大的差别,要怎样证明别人的词典抄袭、盗取你的内容?

要是 esquivallence 这个词出现在《新牛津美语词典》以外的词典里,就一定牵涉到抄袭、盗取,这个词就是为了找出抄袭、盗取而放在那里埋伏的。

当前全球出版业都不景气,特别是在纸书出版领域。中国出版业尤甚,凉意十足。尽管各路专家给出了不同的原因分析,但只有一位专家给出的答案令业内所信服,那就是优质内容的缺失。说到底出版是一个内容为王的产业,没有好的内容,一切都是无本之源。

有位作家说:平庸是这个时代的危险所在,它无法再吸收传统知识;现代生活杂乱无章,令人湮没无闻。一切都掉在浅水中,没有什么沉入深深的井中;一切都是飞短流长,一切都是流言蜚语。

我们应该敢于承认一个基本事实,这个事实便是——在这个平庸的时代,最坏的都活下来了,最好的死去了,我们这些还能逃生的,发挥不出真正的价值。那么,在这个平庸的时代,我们还能做什么呢?

由衷感谢爱思唯尔(Elsevier)公司于2006年6月出版的这套《Encyclopedia of Mathematical Physics》(《数学物理大百科全书》),这是一部不平凡的全面介绍数学物理知识的百科全书。

本书的三位主编(法国巴黎居里大学 Jean-Pierre Francoise 教授、美国费城德雷塞尔大学 Gregory L. Naber 教授和英国牛津大学 Tsou Sheung Tsun 博士)都是长期从事数学物理方面研究的知名学者,他们邀请了包括诺贝尔物理学奖获得者杨振宁教授和英国牛津大学 Roger Penrose 教授在内的34位著名物理学家和数学家,作为本书的编辑顾问委员会成员,组织来自30个国家的439位在物理学和数学相关研究领域做出杰出贡献的理论物理学家和数学家,撰写了400多篇图文并茂的综述性文章。

《数学物理大百科全书》是经长达4年完成的一部内容全面系统、领域涵盖广泛的百科全书,全书特色鲜明,既体现了学科的基础性、独立性、完整性,又注重学科的前沿性、交叉性、应用性,是当今数学物理研究领域最新和最全的百科全书。

本书内容涉及物理学和数学的几乎各个重要研究领域,遍及从经典力学到量子力学、经典场论到量子场论、共形场论到拓扑场论、流体动力学到动力系统、可积系统到无序系统、粒子物理到天体宇宙学、相对论到量子引力、规范理论到统一理论、平衡态统计到非平衡态统计、凝聚态物质到量子信息、变分技术到代数方法、泛函分析到算子代数、路径积分到随机方法、李群到量子群、微分几何到代数拓扑、低维几何到非交换几何、复几何到辛几何等核心领域和方向,本书还特别注重数学物理的最新研究成果和在各领域的最新应用,并提供了大量必要的和重要的参考文献。

本书相比一般的百科全书有一个明显的亮点是它的综述,它可以告诉你你想知道的某个专题的一切。中国科学院院士赫柏林曾留学于哈尔科夫大学,据他回忆当时的考试是由数学物理教授 A. Ya. Povzner 主持,他出的题目是:“把从你生下来以后所知道的贝塞尔函数的一切都告诉我。”据他的学生说:他写了一大摞纸,密密麻麻,然后告诉 Povzner“这是我知道的关于贝塞尔函数知识的提纲,若是需要,我可以展开每一项的具体内容。”于是考试通过。

正如 Gerard't Hooft 所指出的那样:

数学物理这个交叉学科是非常难懂的。百科全书中的某些题目纯粹是物理的,高 T_c 超导电性、破坏水波和磁水动力是完全物理的题目,其中的实验数据比任何高深理论都具有决定性,然而,上调理论、Donaldson-Witten 理论和 AdS/CFT 对应是纯数学的例子。

在编辑中,大量不同作者的短小文章不可避免地被做了适当的变动,在这本百科全书中,理论物理学家和数学家为高等数学物理中的许多重要条目做了简单明了的阐述,所有的文章都包含了供进一步阅读的参考文献,我们盼望这些努力会取得很好的效果。

本书的编者认为:

与狭义的数学和物理学的古老历史相比,数学物理是一门相对较新的独立学科,数学物理国际协会成立于1976年,当然,从古时候起数学与物理学就相互影响,但近几十年来,可能因为我们正身在其中,它们出现了巨大的进展,新的结果和观点以令人目眩的节奏诞生,以至于需要有一本百科全书来搜集整理这些知识。

数学物理把数学和物理学这两个大学科的优势集中到一起,它们的关系是共同发展,一方面,它运用数学这一工具把不断增长的精确性和复杂性这些物理概念组织了起来;另一方面,物理学家为数学家们提供了灵感的源泉,两者关系的经典例子是爱因斯坦的相对论,其中微分几何在物理理论的公式化方面起到了实质性的作用,而物理学相继提出的问题推动了微分几何的发展,巧合的是,当我们在为《数学物理大百科全书》写序言时,正值爱因斯坦创造奇迹100周年。

再三考虑到写这部《数学物理大百科全书》是一个艰巨的项目,如果不是坚信这是一项很有意义的、受益于社会的项目,而且我们会得到众多的支持,那么我们绝不会接受这个任务,我们确实获得了许多支持,包括建议、鼓励和有实用性的帮助,这些支持来自编辑顾问委员会成员和我们的作者,还有其他慷慨地抽时间帮我们完善这本百科全书的人。

数学物理是一门较新的学科,它还没有被清晰地刻画,不同的人对它有不同的理解,在我们选择的题目中,一部分遵循了近期数学物理国际大会的纲要,但主要参照编辑顾问委员会和作者的提议,由于时间和空间的限制,以及我们自身的水平所限,更改了某些冗长的题目,但我们尽量收录了我们认为是核心的课题,尽量覆盖更多的最活跃的领域。

近年在中国对本书的原出版商还是有些负面新闻的,起源是在美国一个名为“知识的代价”网站上,已有全球12 196位科学家签名抵制这家世界上最大的出版商,有人用“学术之春”形容这场运动。

吹响号角的是大名鼎鼎的英国数学家威廉·提摩西·高尔斯(William Timothy Gowers),这位来自剑桥大学的菲尔兹奖得主曾发表了一篇博客文章,号召同行行动起来,抵制世界上最大的出版商爱思唯尔集团。

读到这篇博文的泰勒·内伦(Tyler Neylon)——一位目前在硅谷开公司的数学博士当即给高尔斯教授留了言,第二天,他建立了一个网站,命名为“知识的代价”。

泰勒事后回忆,自己读到那篇博文,就意识到可以做点什么,在他看来,高尔斯是一位拥有号召力的“超级明星”。

迄今为止,数万名科学家在泰勒的网站上签了名,他们发誓,不在爱思唯尔旗下的期刊发论文,不做审稿人,或者不担

任编辑。

尽管如此，我们还是选择了与爱思唯尔的合作，因为一套好的大百科太难得了。

旅法钢琴大师白建宇(Kun-Woo Paik)对钢琴的要求非常苛刻，他在一次与台湾出版人郝明义先生的谈话时说，弹琴弹到现在，职业演奏生涯超过半个世纪，所遇到满意的琴竟不超过5架，如此答案，令见多识广的郝先生也大吃一惊。

在数理方面，近年来国内引进的好的大百科也绝不会超过5部，前苏联五卷本的《数学大百科全书》算一部，日本岩波的《数学百科全书》算一部，总之是屈指可数。

其实这个项目并不是爱思唯尔创始的，据介绍，这个项目开始于 Academic Press，后来由爱思唯尔接手。他们热情的工作人员，把过渡工作做得天衣无缝，并且令人感动的是，相当一部分作者慷慨地把他们的酬劳捐赠给欧洲数学会的发展中国家委员会，我们应该感谢他们为发展中国家所做的一切。

至于我们最关心的问题：谁会去购买这样一套大书，我们充满乐观。大千世界无奇不有，各种购买方式都可能出现。前一阵子，有关霍金打赌输掉关于“上帝粒子”存在性的赌约报道很多。

实验证明霍金输掉了这场赌约，霍金坦承自己输得心服口服并祝愿希格斯获得诺贝尔奖。希格斯透露，在宣布发现新粒子后，霍金曾与他联系并表示支票已寄出，希格斯说，“他不仅是给我一个人钱。我想他还会寄100美元给密歇根大学的戈登·凯恩。”

这场赌约的另一位赢家凯恩对来自霍金的美元欣然接受。“我坚信希格斯玻色子一定会被找到。发现希格斯玻色子真是太棒了，它证实了长久以来的猜想，进一步加强了粒子物理‘标准模型’的事实根据。打赌获胜是锦上添花。”凯恩表示要把赢来的钱花在刀刃上，所有的钱都要用于搞研究。

霍金可能已经习惯了以输掉赌约的方式推进科学的普及。

1975年，霍金曾关于天蝎座X-1是否包含黑洞打赌，后来认输，为赢家订阅了1年的《阁楼》杂志。

1991年，霍金又与人赌上了，这次赌的是裸奇点是否存在，霍金再次输了。

第三次打赌发生在1997年，霍金同美国物理学家约翰·普雷斯基尔打赌，认为黑洞部不会摧毁它们吞噬的一切信息，霍金于2004年7月21日当众表示输掉了这场赌约，并送给普雷斯基尔一套板球百科全书。

关于希格斯玻色粒子的赌约则是他的第四场赌约。这30多年来，霍金通过杂志、书籍和一点点美元，让更多的人了解到这些科学最前沿的问题。在100美元的赌约背后，希格斯的远见和霍金的牺牲精神都值得称道。

我们期待下一个赌约会以这样一套百科全书来结束。

著名力学家周培源90岁生日时，北京大学全体师生用“献身科学，教育英才；功在国家，造福将来；寿齐嵩岱，德被春荑；祝嘏欢呼，消欤盛哉”的贺词赞扬他们的老校长。斗胆借用一下，庆祝这套书在中国的出版，当不为过。

刘培杰

2015年11月1日

于哈工大